IUP LIBRARIES
A000001964813

Nussbaum

Semiconductor Device Physics

7-CK
N8

PRENTICE-HALL EE SERIES

Nussbaum

Semiconductor Device Physics

INTERNATIONAL SERIES IN ELECTRICAL ENGINEERING

051

PHYSICAL CONSTANTS

Name	Symbol	Magnitude
Planck's constant	h	6.63×10^{-34} joule-sec
Magnitude of the charge on the electron	e	1.60×10^{-19} coul
Mass of a free electron	m	9.11×10^{-31} kgm
Boltzmann's constant	k	1.38×10^{-23} joule/deg, or 8.62×10^{-5} eV/deg
Permittivity of free space	$\mathcal{E}_0$	8.85×10^{-12} farad/m
Avogadro's constant	L	6.02×10^{26} molecules/ kgm-molecular wt.

SEMICONDUCTOR DEVICE PHYSICS

PRENTICE-HALL INTERNATIONAL SERIES IN ELECTRICAL ENGINEERING

BLACKWELL AND KOTZEBUE *Semiconductor-Diode Parametric Amplifier*
NUSSBAUM *Semiconductor Device Physics*

PRENTICE-HALL, INC.
PRENTICE-HALL INTERNATIONAL, INC., UNITED KINGDOM AND EIRE
PRENTICE-HALL OF CANADA LTD., CANADA
BERLINER UNION, WEST GERMANY AND AUSTRIA

ALLEN NUSSBAUM
American Electronics Laboratories
Lansdale, Pennsylvania

SEMICONDUCTOR DEVICE PHYSICS

PRENTICE-HALL, INC.
Englewood Cliffs, N.J./1962

Prentice-Hall Electrical Engineering Series
WILLIAM L. EVERITT, *editor*

SERIES IN SOLID STATE PHYSICAL ELECTRONICS
NICK HOLONYAK, JR., *editor*

©1962 by PRENTICE-HALL, INC., Englewood Cliffs, N. J. All rights reserved. No part of this book may be reproduced in any form, by mimeograph or any other means, without permission in writing from the publishers.

Library of Congress Catalog Card Number 62-7844

Printed in the United States of America
80612C

Library
State College
Indiana, Penna.

537.62 N944
C.1

to my parents, Sarah and Harry

PREFACE

The Objectives of This Book

The study of the physics of semiconductors bridges a wide gap. It is based on fundamentals such as quantum theory, statistical mechanics, and irreversible thermodynamics, and it applies to devices like transistors and thermoelectric generators. The purpose of this book is to help the science student and the engineer in industry understand the connection between these fundamental concepts and the properties of the semiconductors and devices with which he is (or will be) dealing. The material presented here was originally designed for a group of research assistants in semiconductor physics and transistor development. It has been subsequently modified so that it will be more suitable as a text at the intermediate undergraduate to beginning graduate level or as a self-study book for people with a background in science or engineering.

The Qualifications Expected of the Reader

The user of this book is required to have an acquaintance with determinants and with differentiation and integration, plus a brief introduction to differential equations, and the equivalent of a first-year college physics course including a non-mathematical introduction to atomic physics. Vector and tensor analysis are used where needed, and these topics are covered in the appendix. Special topics, such as the method of Lagrange's multipliers, are explained at the appropriate places in the text. Since equations, like pictures, are worth many words, a large amount of mathematics is used in this book.

The Method of Presentation

In order to give the user an intuitive understanding of the physical principles of semiconductor physics and devices, many analogies and mechanistic explanations are used. While this approach lacks the rigor of the standard monographs in quantum and statistical physics, it does help the beginner in becoming familiar with a fairly complex field. The problems are an integral part of the text and should be worked as they are encountered. Some are designed to give familiarity with the magnitudes of parameters, some show the applications of various types of mathematical techniques, and others extend the principles presented.

For those who use this book as a text, it should be noted that the first four chapters constitute an introduction to the theory of diodes and transistors, while the remaining four chapters go on to a more thorough study of the basic properties of semiconductor materials.

Acknowledgments

The preparation of this book owes much to other works which are listed as references. I would like to thank a former teacher, Prof. H. B. Callen of the University of Pennsylvania, whose book serves as a basis for Chapter II and whose superb course on mathematical physics furnished many of the ideas used in the appendix. My principal indebtedness is to my former colleagues at the Honeywell Research Center. In particular, the ideas and remarkable insight of Dr. Charles A. Domenicali will be found reflected in many places throughout this book. Dr. John S. Blakemore critically read the manuscript, made many valuable improvements, and provided unpublished data. In addition, he wrote the material on recombination which appears in Chapter IV. I have received a great deal of help and encouragement from Dr. Roy Morrison and Dr. Carl Nomura. Dr. Donald Long provided unpublished data and Dr. Donald Blue prepared a difficult sketch of the crystal structure of tellurium. I wish to thank Dr. Murray Bullis of Texas Instruments, Inc. for the gift of a copy of his thesis, part of which is used in Chapter VII, and Dr. William Peria of the University of Minnesota for pointing out some serious errors. I would like to acknowledge my debt to Dr. Finn J. Larsen, Vice President for Research, and Dr. Van W. Bearinger, Director of Research, who created the unique atmosphere which made this work possible. Finally, I wish to express appreciation to Mrs. Corrine Lehr, who typed and retyped the manuscript many times, and to Mrs. Verna Squier, who did the original drawings.

ALLEN NUSSBAUM

CONTENTS

1

QUANTUM PHYSICS

1-1. INTRODUCTION, 1.
1-2. FUNDAMENTAL QUANTITIES AND THE MKS SYSTEM OF UNITS, 1.
1-3. DERIVED QUANTITIES, 2.
1-4. ELECTROSTATICS AND ELECTROMAGNETISM, 4.
1-5. THE WAVE EQUATION OF CLASSICAL PHYSICS, 8.
1-6. THE RELATIONS OF PLANCK, EINSTEIN, AND DE BROGLIE, 11.
1-7. THE SCHROEDINGER EQUATION, 18.
1-8. THE QUANTUM OSCILLATOR, 21.
1-9. POTENTIAL WELLS, 23.
1-10. THE POTENTIAL BOX, 26.
1-11. THE HYDROGEN ATOM, 29.
1-12. ELECTRON SPIN AND THE EXCLUSION PRINCIPLE, 34.
1-13. ELECTRONIC ENERGY BANDS, 36.

2

THERMOSTATICS AND STATISTICAL PHYSICS

2-1. TEMPERATURE, HEAT, AND INTERNAL ENERGY, 44.
2-2. ENTROPY, 45.
2-3. TEMPERATURE, PRESSURE, AND CHEMICAL POTENTIAL, 46.
2-4. REVERSIBLE AND IRREVERSIBLE PROCESSES; THE MEANING OF ENTROPY, 48.
2-5. ENTHALPY AND THE HELMHOLTZ AND GIBBS FUNCTIONS, 51.
2-6. THE STATISTICAL DISTRIBUTIONS, 55.
2-7. THE PERFECT GAS, 60.
2-8. THE FERMI LEVEL AND THE ELECTROCHEMICAL POTENTIAL, 64.

3

ENERGY BANDS AND CARRIER STATISTICS

3-1. CONDUCTORS, INSULATORS, AND SEMICONDUCTORS, 70. 3-2. CONDUCTION BY HOLES, 78. 3-3. THE STATISTICS OF SEMICONDUCTORS, 80. 3-4. THE PN-JUNCTION, 91.

4

THE THEORY OF JUNCTION DEVICES

4-1. THE PN-JUNCTION IN EQUILIBRIUM, 94. 4-2. THE LOW-LEVEL JUNCTION, 103. 4-3. THE JUNCTION TRANSISTOR, 113. 4-4. THE PNN−-DIODE AND TRANSISTOR, 122. 4-5. JUNCTION DEVICES UNDER HIGH-LEVEL INJECTION, 129. 4-6. GENERATION AND RECOMBINATION PROCESSES IN JUNCTION DEVICES, 137.

5

PERIODIC STRUCTURES AND ENERGY BANDS

5-1. SYMMETRY AND LATTICES, 151. 5-2. POINT GROUPS, GLIDE LINES, AND PLANE GROUPS, 158. 5-3. THE BRAVAIS LATTICES, 163. 5-4. THE THREE-DIMENSIONAL POINT GROUPS, 170. 5-5. THE SPACE GROUPS, 177. 5-6. CRYSTAL STRUCTURE AND BRILLOUIN ZONES, 183. 5-7. THE TIGHT-BINDING OR LCAO METHOD, 190. 5-8. THE BAND STRUCTURE OF TELLURIUM, 198. 5-9. THE ENERGY BANDS OF SILICON AND GERMANIUM, 203.

6

IRREVERSIBLE THERMODYNAMICS

6-1. THE ONSAGER PRINCIPLE, 212. 6-2. CONJUGATE FORCES AND FLUXES, 215. 6-3. THERMOELECTRIC POWER, 218. 6-4. THE PELTIER AND THOMPSON COEFFICIENTS, 220.

7

TRANSPORT THEORY

7-1. AN INTRODUCTION TO THE GALVANOMAGNETIC COEFFICIENTS, 223. 7-2. THE BOLTZMANN TRANSPORT EQUATION, 230. 7-3. THE SOLUTION OF THE BOLTZMANN EQUATION, 231. 7-4. CALCULATION OF THE ELECTRICAL AND HEAT CURRENT DENSITIES, 233. 7-5. SOLUTION OF THE BOLTZMANN EQUATION IN THE PRESENCE OF A MAGNETIC FIELD FOR ELLIPSOIDAL ENERGY SURFACES, 236. 7-6. THE GALVANOMAGNETIC COEFFICIENTS, 240. 7-7. INTERPRETATION OF EXPERIMENTAL DATA FOR N-TYPE GERMANIUM AND SILICON, 242. 7-8. THE SOLUTION TO THE BOLTZMANN EQUATION FOR WARPED ENERGY SURFACES, 253. 7-9. THE ELECTRICAL PROPERTIES OF TELLURIUM, 262. 7-10. THE THERMOMAGNETIC EFFECTS, 273.

8

SEMICONDUCTOR PARAMETERS

8-1. INTRODUCTION, 277. 8-2. RESISTIVITY AND COMPENSATION, 277. 8-3. DETERMINATION OF IMPURITY CONCENTRATIONS, 283. 8-4. APPLICATIONS IN THERMOELECTRICITY, 287. 8-5. SCATTERING AND MOBILITY, 296. 8-6. DETERMINATION OF MINORITY CARRIER LIFETIME, 308.

APPENDIX

A-1. TAYLOR SERIES, 313. A-2. VECTOR ALGEBRA, 315. A-3. VECTOR DIFFERENTIAL OPERATORS, 320. A-4. COORDINATE TRANSFORMATIONS AND MATRICES, 328. A-5. INTRODUCTION TO TENSORS, 331.

INDEX 335

1

QUANTUM PHYSICS

1-1 Introduction

The purpose of this chapter is the study of the motion of electrons in atoms and crystals. This forms part of the subject known as *quantum mechanics* and we shall present its ideas in a plausible, rather than a rigorous, way. Before taking up this topic, however, we shall review the concepts and units of classical physics that are used in this book.

1-2 Fundamental Quantities and the MKS System of Units

All quantities used in physics can be expressed in terms of the following *fundamental* ones: *mass* m, *length* l, *time* t, *charge* q, and *temperature* T. These five are considered to be undefinable and we think of them in an intuitive fashion. For example, we associate with temperature the concepts of "hotness" or "coldness" without trying to give a rigorous definition of these terms.

The study of the motion of particles or bodies uses the first three of these fundamental quantities. The units of measurement which form the basis of the metric system are the meter (m) for length, the kilogram (kg) for mass, and the second (sec) for time. This choice of units is called the *MKS system* and will be the primary system used in this book. Other convenient units for length we shall use are the micron (μ), $1\ \mu = 10^{-6}$m, and the Angstrom unit (A), $1\ \text{A} = 10^{-10}$m; for time, we also use the microsecond (μsec), $1\ \mu\text{sec} = 10^{-6}$ sec.

1-3 Derived Quantities

We can now define many other familiar physical concepts, called *derived quantities*, in terms of the five fundamental ones of Section 1-2. Some of the most frequently used are shown in Table 1-1.

TABLE 1-1

Derived quantity	*Symbol*	*Definition*	*Unit*	
Speed or velocity*	v	$v = \dfrac{dl}{dt}$	m/sec	(1-1)
Acceleration	a	$a = \dfrac{dv}{dt} = \dfrac{d^2l}{dt^2}$	m/sec²	(1-2)
Momentum	p	$p = mv$	kg-m/sec	(1-3)
Force	F	$F = \dfrac{dp}{dt}$	newton	(1-4)
Work†	W	$W = \int \mathbf{F}\cdot d\mathbf{l}$	joule	(1-5)
Power	$\mathcal{P}$	$\mathcal{P} = \dfrac{dW}{dt}$	watt	(1-6)

*Velocity is defined as speed in a given direction.

†Vectors are denoted by bold-face type. An introduction to vector algebra covering the applications made in this chapter will be found in Appendix A-2.

There are a few derived quantities which should be briefly discussed. One of these is *energy E*, defined as the ability of a body to do work. The unit of energy is the joule, as for work. The energy which a body possesses by virtue of its position is called *potential energy V*; that which is due to its motion is *kinetic energy T*. We can find a general formula for kinetic energy from Eqs. (1-4) and (1-5). Let a force F along the x-axis impart a speed v to a body originally at rest. Then the work done is

$$W = \int F\,dx = \int \frac{dp}{dt}\,dx \tag{1-7}$$

For a constant mass, by Eq. (1-3),

$$W = m\int \frac{dv}{dt}\,dx = m\int_0^v v\,dv = \tfrac{1}{2}mv^2$$

Thus, the amount of work done in producing the motion, which by definition is the kinetic energy, is

$$T = \tfrac{1}{2}mv^2 \tag{1-8}$$

Another quantity we shall discuss briefly is *pressure P*, which is defined in terms of the force F exerted over an area A as

$$P = \frac{dF}{dA} \tag{1-9}$$

The MKS unit of pressure is thus the newton/m². A more common unit is the atmosphere (atm), and 1 atm is approximately 10^5 newton/m².

To find F from Eq. (1-9), we integrate

$$F = \int P\,dA \tag{1-10}$$

Substituting Eq. (1-10) into Eq. (1-5) yields

$$W = \iint P\,d\mathbf{A}\cdot d\mathbf{l}$$

where $d\mathbf{A}$ is a vector whose magnitude is dA and whose direction is along the normal to dA. Now $d\mathbf{A}\cdot d\mathbf{l}$ represents the volume dV which is generated by the area dA undergoing a displacement $d\mathbf{l}$. Hence

$$W = \int P\,dV$$

and

$$dW = P\,dV \tag{1-11}$$

There is a group of derived quantities associated with rotational motion. An angular displacement θ will usually be measured in radians, the radian

TABLE 1-2

Derived quantity	*Symbol*	*Definition*	*Unit*
Angular Velocity	ω	$\omega = \dfrac{d\theta}{dt}$	sec⁻¹
Angular momentum	p_θ	$\mathbf{p}_\theta = \mathbf{r}\times\mathbf{p}$	kg-m²/sec

being a dimensionless quantity which gives the ratio of the arc-length subtended by the displacement to the corresponding radius. Other angular quantities we shall use are shown in Table 1-2.

1-4 Electrostatics and Electromagnetism

The ideas covered so far in this chapter can be extended to the study of stationary and moving electric charges by the introduction of charge as the fourth fundamental quantity. The unit of charge is called the *coulomb* and the set of units that we now have established is called the *MKSC* or *absolute practical* system. We will define *quantity of charge* q as that property of two charged particles to which the mutual force at a given distance in a given medium is directly proportional, or

$$F \propto q_1 q_2$$

where q_1 and q_2 are the respective charges on each particle. Coulomb showed experimentally that for two given charges in a given medium separated by a distance r

$$F \propto 1/r^2$$

Define the *permittivity* ϵ of a medium as that property of the medium to which the electrostatic force at a given distance is inversely proportional. Combining these relations yields

$$F = \frac{C\, q_1 q_2}{\epsilon r^2}$$

where C is a proportionality constant. It is convenient to let $C = 1/(4\pi)$, since this simplifies many of the equations of advanced electromagnetism, and

$$F = \frac{q_1 q_2}{4\pi\epsilon r^2} \tag{1-12}$$

This equation is known as *Coulomb's law.* A convenient natural standard of charge is the electron; this charge is assigned the value of -1.6×10^{-19} coulomb* to make it consistent with the international ampere. We shall use the symbol

$$-e = -1.6 \times 10^{-19} \text{ coulomb} \tag{1-13}$$

for this quantity. It is found experimentally that the force between two

---

*For purposes of simplifying numerical examples, we shall carry only a few significant figures in the values of physical constants.

one-coulomb charges one meter apart in free space is about 9×10^9 newtons. By Eq. (1-12), the value of ϵ_0, the permittivity of free space, is

$$1/(36\pi \times 10^9) = 8.85 \times 10^{-12} \text{ coulomb}^2 \text{ sec}^2/\text{kg-m}^3$$

The ratio ϵ/ϵ_0 is called the *relative permittivity* or *dielectric constant* ϵ_r.

The space surrounding an electric charge, in which forces are exerted on other charges, is called the *field* of the charge. The *field intensity* $\mathcal{E}$ is the force per unit charge

$$\mathcal{E} = F/q \tag{1-14}$$

Because of these forces, the motion of a charge in a field implies the performance of work. The work per unit charge in such a process is defined as the *electrostatic potential difference* $\mathcal{V}$ between the initial and final positions of the charge, or

$$\mathcal{V} = W/q \tag{1-15}$$

The sign of $\mathcal{V}$ is taken as positive when work is done *on* a charge. By Eqs. (1-7) and (1-14),

$$\mathcal{V} = -\int (F/q)\, dx = -\int \mathcal{E}\, dx \tag{1-16}$$

where the negative sign is used to show that $\mathcal{V}$ will increase if the direction of motion is opposed to the direction of the force. The unit of $\mathcal{V}$ is called the *volt,* and it is simpler to refer to $\mathcal{V}$ as the *voltage.* By Eq. (1-16) we see that the unit of $\mathcal{E}$ must then be the volt/meter.

The energies which electrons in solids possess are too small to be conveniently measured in joules, and we use a unit called the *electron-volt* (eV), which is defined as the work required to move an electron through a potential difference of 1 volt. This work is

$$W = e\mathcal{E} = 1 \text{ eV} = (1.6 \times 10^{-19} \text{ coulomb})\ (1 \text{ volt})$$

or

$$1 \text{ eV} = 1.6 \times 10^{-19} \text{ joule} \tag{1-17}$$

A *capacitor* consists of two charged bodies, holding equal but opposite charges, and separated by an insulating medium. The *capacitance* C of such a device is defined as the ratio of the positive charge to the difference in potential, or

$$C = q/\mathcal{V} \tag{1-18}$$

The unit of capacity is the *farad.*

PROBLEM 1-1 Show that the units of ϵ and ϵ_0 are farads/meter.

When charges move in a conductor *current* is said to flow. Current I is quantitatively defined as

$$I = dq/dt \tag{1-19}$$

The unit of current is the *ampere.*

Current density J is the current which flows through a unit area perpendicular to the direction of flow. The unit is thus the ampere/meter^2. The motion of the charges is caused by establishing a difference in potential $\mathcal{V}$ and the ratio $\mathcal{V}/I$ is called the *resistance* R, the unit being the *ohm.* Thus

$$\mathcal{V} = IR \tag{1-20}$$

The resistance of a unit cube of material is called the *resistivity* or *specific resistance* ρ, and the reciprocal of ρ is called the *conductivity* σ.

$$\sigma = 1/\rho$$

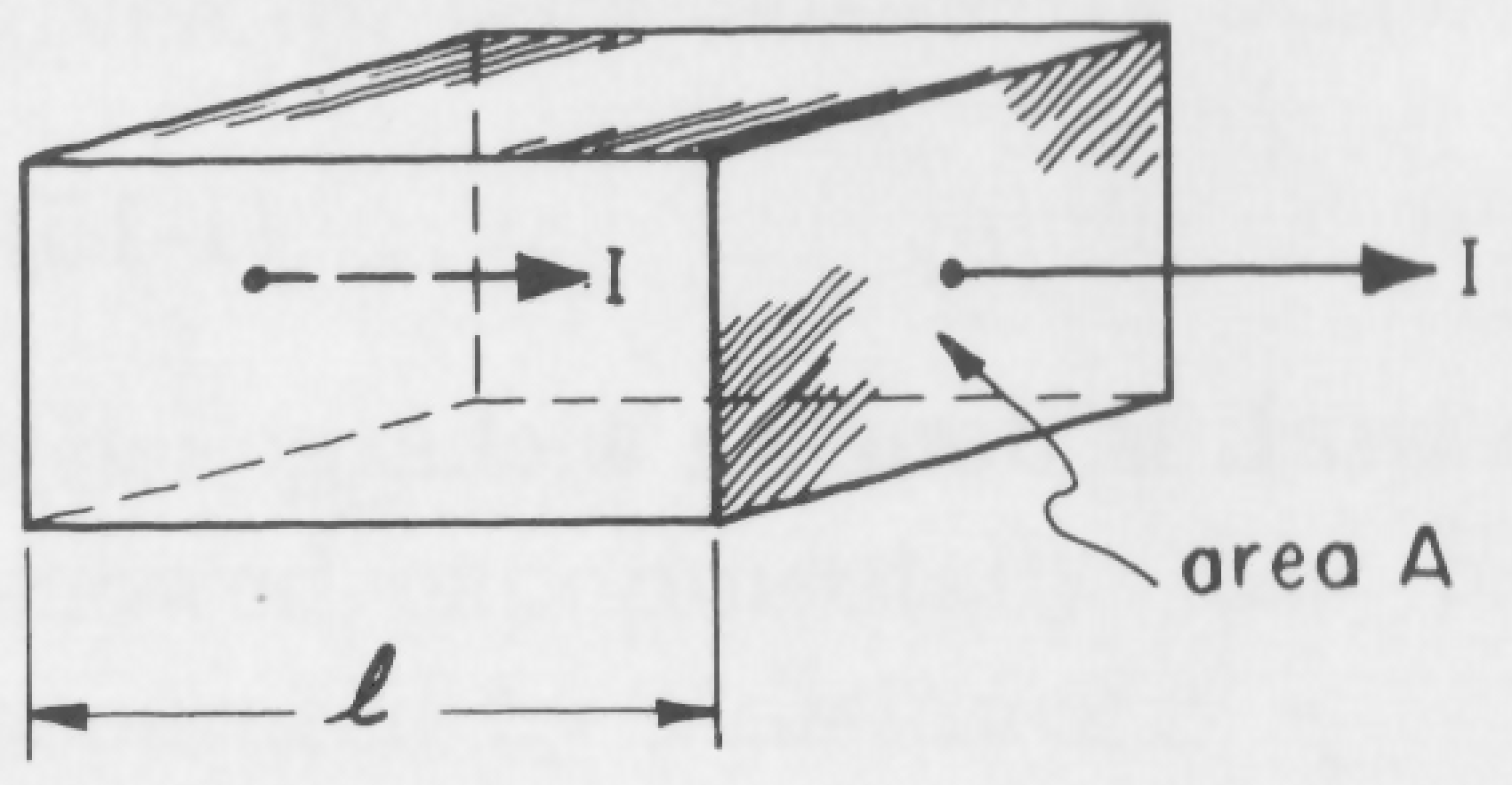


FIG. 1-1. A conducting block used to show the relation between current I and current density J.

Suppose we apply a potential $\mathcal{V}$ across two opposite faces of a rectangular block of a material and produce a current I, as in Fig. 1-1. It is found experimentally that the resistance of such a block is directly proportional to the length l and inversely proportional to the area A, so that

$$R = \rho l/A = l/\sigma A$$

where the proportionality constant is taken as ρ. Note that when $l = 1$ and $A = 1$ for a unit cube, then $R = \rho$, as the definition requires. By the definition of current density J,

$$J = I/A$$

Putting these two relations into Eq. (1-20) yields

$$\mathcal{V} = Jl/\sigma$$

But the electric field in the block is related to $\mathcal{V}$ by Eq. (1-16).

$$\mathcal{V} = -\int_l^0 \mathcal{E}\,dx = \mathcal{E}l$$

where we have integrated from l to 0, since this is the direction in which work would be done *on* a charge. Then we have

$$J = \sigma\mathcal{E} \tag{1-21}$$

PROBLEM 1-2 Let the block of Fig. 1-1 contain n electrons per unit volume, each one moving with a speed v in the direction shown. Prove that the current density is given by

$$J = nev \tag{1-22}$$

The *mobility* μ of an electron in a conductor is the property which indicates how fast it will move when an electric field of a given strength is applied. It is defined as the velocity per unit applied field

$$\mu = v/\mathcal{E} \quad m^2/\text{volt-sec} \tag{1-23}$$

and, by Eqs. (1-21) and (1-22),

$$\sigma = ne\mu \tag{1-24}$$

The power associated with an electric current resulting from a constant voltage can be obtained from Eq. (1-15) by using Eqs. (1-6) and (1-16)

$$\mathcal{P} = dW/dt = \mathcal{V}dq/dt = \mathcal{V}I \tag{1-25}$$

Thus a watt is equivalent to a volt-ampere or a joule/second.

To introduce the electromagnetic concepts we shall need in later chapters, consider a stream of electrons all moving with speed v at right angles to a magnetic field, as shown in Fig. 1-2. The electron stream constitutes a current. It is known that a magnetic field will exert a force on a conductor carrying a current. Hence there will be a force on the electrons due to the magnetic field. It is found experimentally that this force is perpendicular to both the direction of motion and the direction of the field, so that all three are mutually at right angles, as shown in the figure. We would expect the force to be proportional to v, and to a property of the magnetic field which we call the *flux-density* or *field intensity* B, so that quantitatively we define B by the equation

$$\mathbf{F} = e(\mathbf{v} \times \mathbf{B}) \tag{1-26}$$

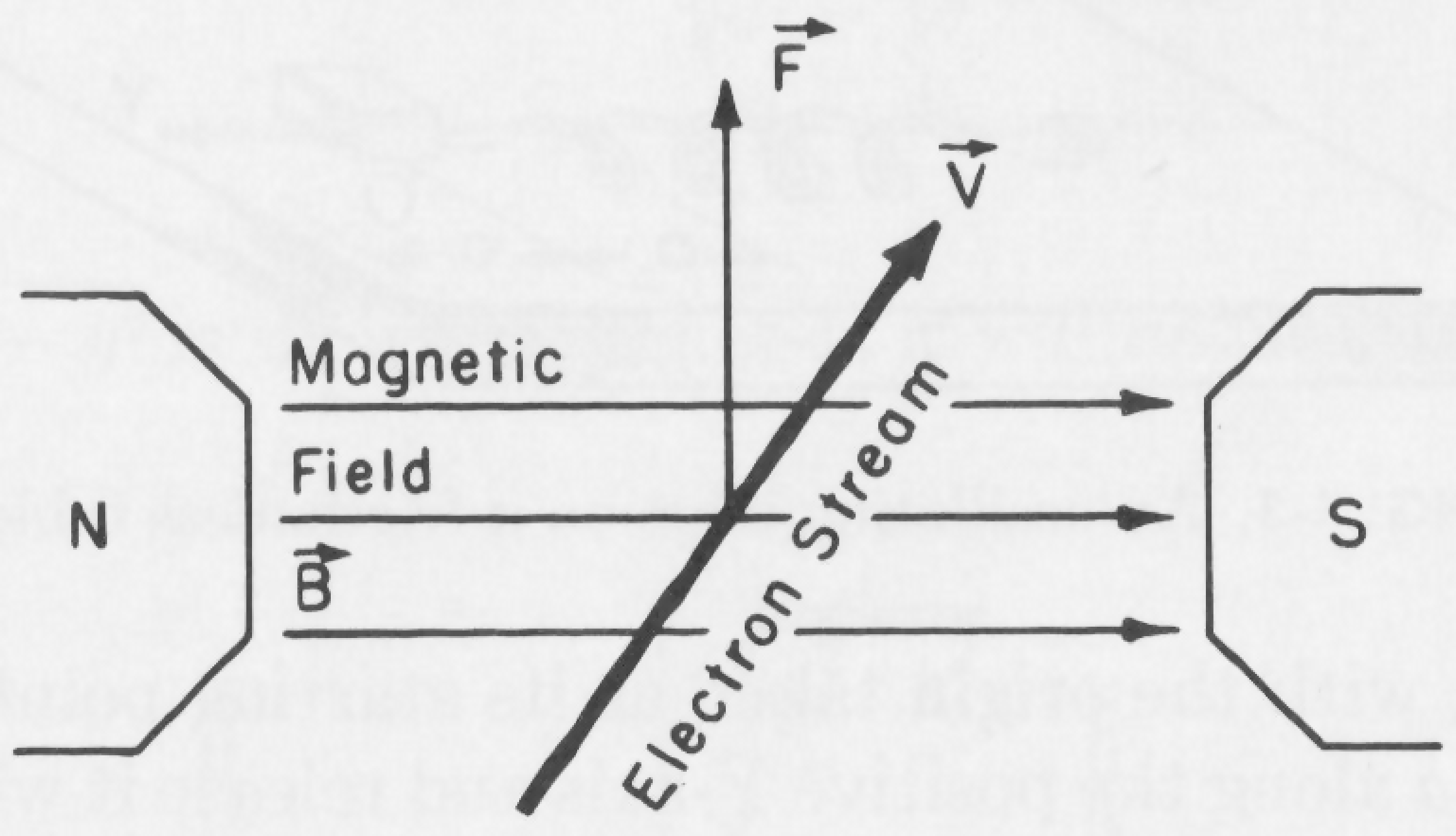


FIG. 1-2. Geometrical relation between electron velocity **v**, magnetic field **B**, and force **F**.

The units of B in the MKS-system are the *weber/meter*^{2}. An older unit for B is the *gauss;* this unit is more commonly used in reporting experimental results. The two are related by

$$1 \text{ weber}/m^2 = 10^4 \text{ gauss}$$

1-5 The Wave Equation of Classical Physics

A *harmonic force* is defined as a force whose magnitude is proportional to the distance from the point of application to a fixed point, called the force-center, and whose direction is such that the force always points towards this center. For example, Hooke's law tells us that the restoring force of a spring is proportional to the elongation but opposite in direction, so that a spring exerts a harmonic force. Quantitatively,

$$F = -\kappa y \tag{1-27}$$

where y is the distance from the force-center to the point of application, κ is called the *Hooke's law constant,* and the minus sign expresses the fact that the elongation and restoring force are opposite in direction. Suppose we have a spring lying on a frictionless table with one end fastened down and a mass m attached to the other end, as shown in Fig. 1-3. Let the mass move along OY, with the origin taken as its starting point. If we pull the mass a distance a along the positive Y-axis and release it with zero velocity, it will vibrate back and forth, alternately compressing and extending the spring. Using the definition of force Eq. (1-4), Eq. (1-27) becomes

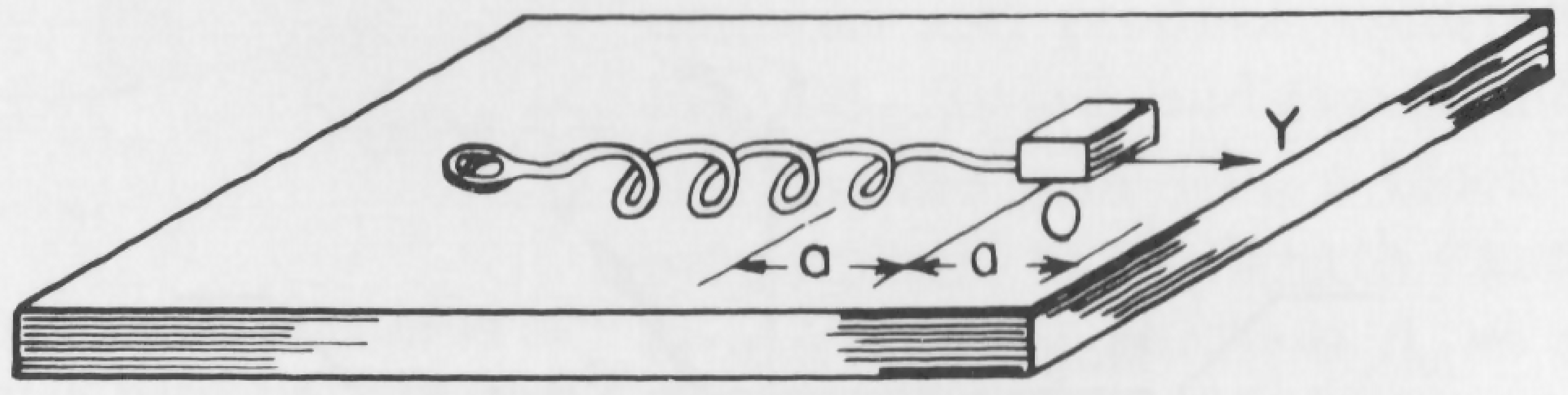


FIG. 1-3. An oscillating mass on a frictionless table.

$$\frac{d^2y}{dt^2} + \omega^2 y = 0 \tag{1-28}$$

where $$\omega^2 = \kappa/m$$

Equation (1-28) is a differential equation with a solution.

$$y = A \cos(\omega t + B) \tag{1-29}$$

PROBLEM 1-3

(a) Verify that Eq. (1-29) satisfies Eq. (1-28).

(b) Show that the solution, Eq. (1-28), subject to the boundary conditions $y = a$ and $v = 0$ at $t = 0$, is

$$y = a \cos \omega t \tag{1-30}$$

Find the corresponding solution for $y = a$ and $v = v_0$ at $t = 0$.

(c) The quantity B in the solution is called the *phase.* What is its physical meaning?

(d) Show that ω, the angular velocity, is related to the frequency of vibration ν by

$$\nu = \omega/2\pi \tag{1-31}$$

PROBLEM 1-4

(a) Show that the probability dP of finding the mass m in a small interval dy located a distance y from the origin is

$$dP \propto \frac{dy}{v}$$

(b) Find v from Eq. (1-30) and show that

$$dP \propto \frac{dy}{\sqrt{a^2 - y^2}} \tag{1-32}$$

A plot of $1/\sqrt{a^2 - y^2}$ is shown in Fig. 1-4; it will be used later.

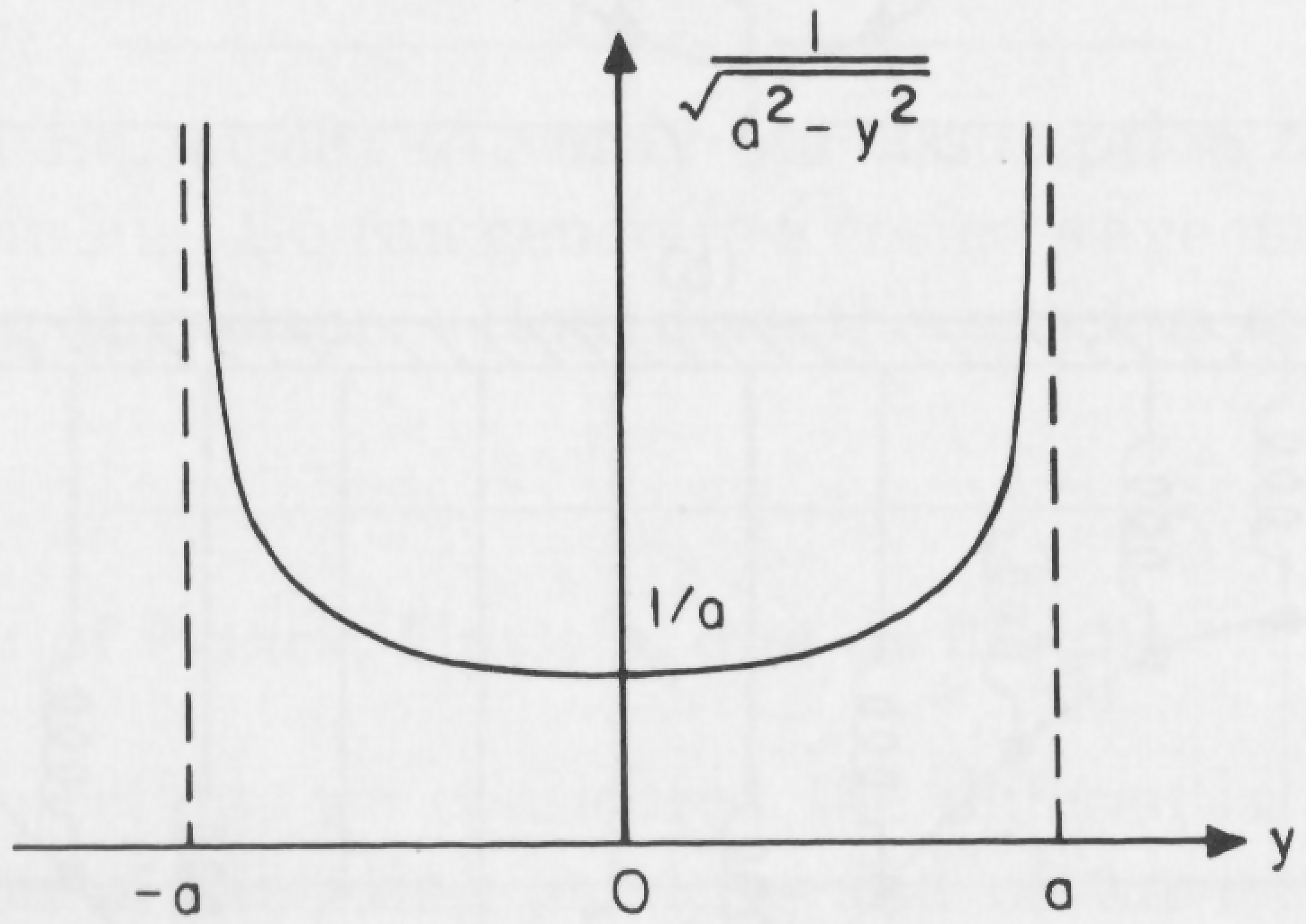


FIG. 1-4. The shape of the position-probability curve for the harmonic oscillator. (From V. Rojansky, *Introductory Quantum Mechanics,* Prentice-Hall, New York, 1938, with permission.)

Now consider a group of identical springs and masses, with the masses arranged to lie on the curve

$$y = a \cos kx \tag{1-33}$$

and vibrating parallel to the y-axis, as shown in Fig. 1-5a. The quantity k, called the *propagation constant,* can be expressed in terms of the *wave* *lenght* λ, defined as the length of one cycle of cos kx. When $x = 0$ or λ, then $y = a$; for $x = \lambda$,

$$k\lambda = 2\pi \tag{1-34}$$

or

$$k = 2\pi/\lambda$$

and Eq. (1-33) can be written

$$y = a \cos (2\pi x/\lambda) \tag{1-35}$$

Let us now assume that the masses can all oscillate in such a way that the cosine curve can move to the right with a velocity v, as shown in Fig. 1-5b.

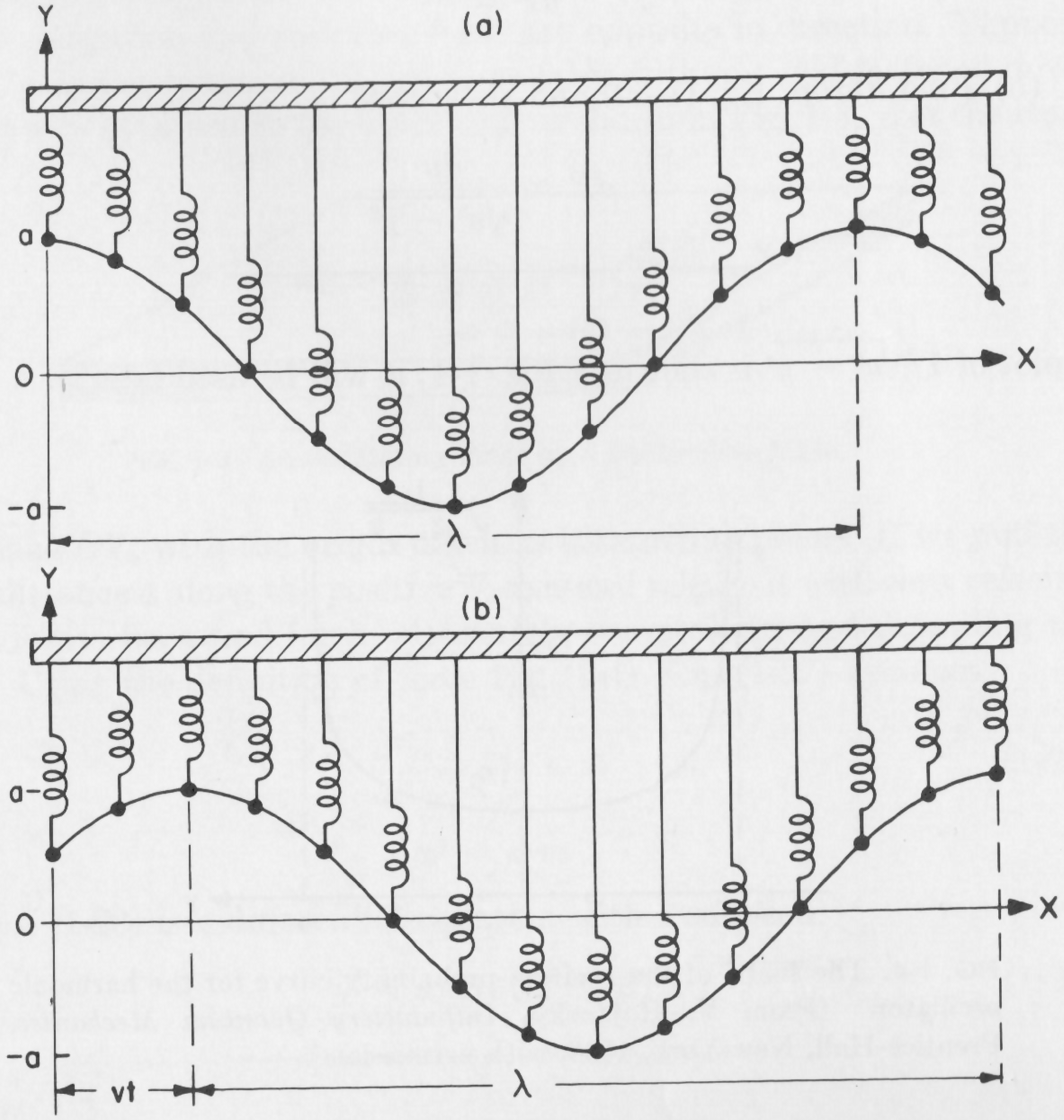


FIG. 1-5. Propagation of a cosine wave by a group of harmonic oscillators.

PROBLEM 1-5 Show that Eq. (1-35) must be modified to

$$y = a \cos [2\pi(x - vt)/\lambda] \tag{1-36}$$

to describe the moving wave of Fig. 1-5b.

Since the velocity of the wave is the product of the number of cycles per second ν and the length of each cycle λ, then

$$v = \lambda\nu \tag{1-37}$$

and Eq. (1-36) can be expressed as

$$y = a \cos (kx - \omega t) \tag{1-38}$$

PROBLEM 1-6 By partial differentiation of Eq. (1-38) and the use of Eqs. (1-34) and (1-37), show that

$$\frac{\partial^2 y}{\partial t^2} = v^2 \frac{\partial^2 y}{\partial x^2} \tag{1-39}$$

Equation (1-39), which is the equation satisfied by a wave of the form of Eq. (1-38), is generally known as the *classical wave equation.*

PROBLEM 1-7 Use Eq. (1-38) to verify our assumption that the group of vibrating masses in Fig. 1-5 can generate a cosine wave moving to the right without distorting the shape. How does the concept of phase enter this problem?

1-6 The Relations of Planck, Einstein, and de Broglie

In the preceding section, we considered the differential equations which govern the motion of vibrating particles and of waves formed by these particles. The particles we are discussing are those of ordinary dimensions: marbles or ping-pong balls, for example. When we get down to atomic sizes, we must modify these classical equations; in order to understand how to do so, we will consider in this section some of the most profound ideas of modern physics.

First, let us look at the *photoelectric effect.* A vacuum tube with a sodium electrode serving as an electron emitter and with a metal collector is connected in the circuit of Fig. 1-6. When ultraviolet light strikes the emitter and the collector is positively charged, all the electrons will be collected, and current will flow through the microammeter. If the voltage on the collector is reduced to zero and then made negative, some of the electrons will be repelled. Eventually, the voltage will reach a *cut-off*

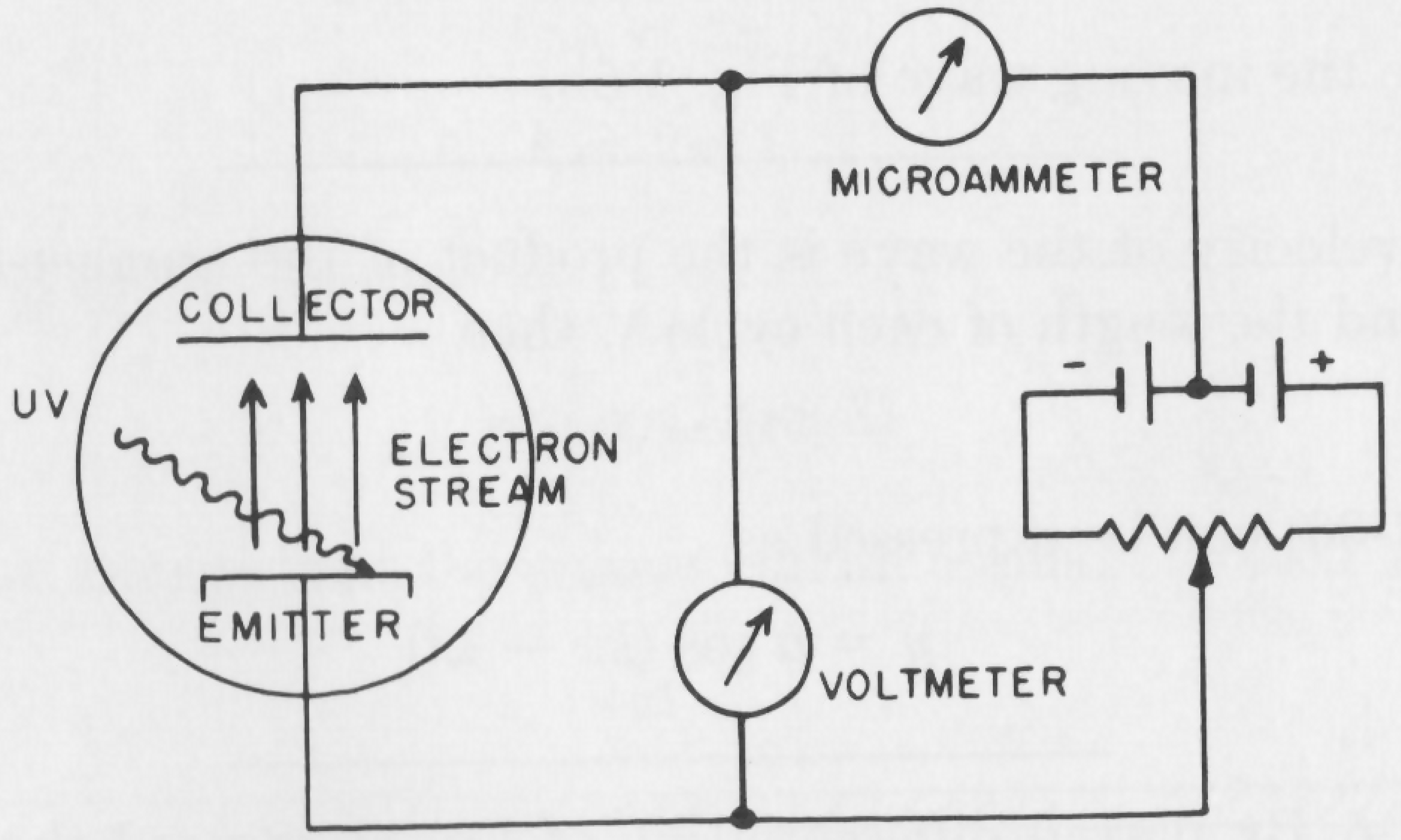


FIG. 1-6. Circuit for studying the photoelectric effect. (From R. L. Sproull, *Modern Physics*, John Wiley, New York, 1956, with permission.)

value $\mathcal{V}_0$, which is needed to just prevent the most energetic electron from reaching the collector. By Eq. (1-15), this maximum energy is $-e\mathcal{V}_0$, where e is the value of the charge on an electron. A typical plot of current vs. voltage is shown in curve A of Fig. 1-7.

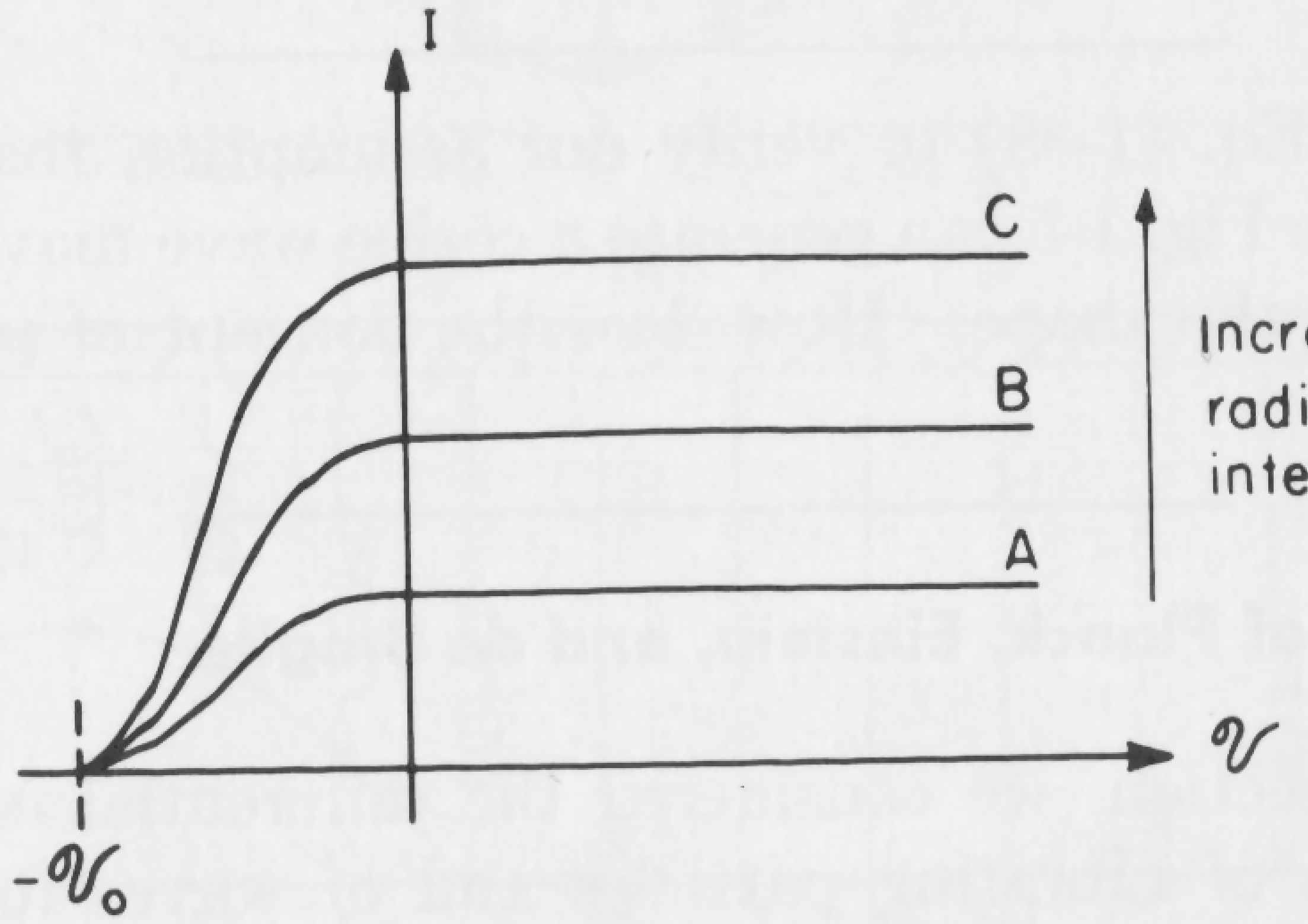


FIG. 1-7. The effect of increasing radiation intensity on the current vs. voltage curves for the arrangement of Fig. 1-6. (From Sproull, *op. cit.*)

Suppose we now make the light stronger by increasing the intensity of radiation (the number of joules/sec striking the emitter.) We would expect this added energy to raise the kinetic energy of the emitted electrons, and thus make $\mathcal{V}_0$ go more negative. Curves B and C of Fig. 1-7 show that this prediction is *not* in agreement with experiment; only the number of electrons, i.e., the current, rises. Next, let us increase the frequency of the light, leaving the intensity fixed. The experimental results of Fig. 1-8 show

that $\mathcal{V}_0$, and thus the energy, *increases* with frequency ν, which is again a surprising result. Further, if we plot $-\mathcal{V}_0$ vs. ν from this graph, we get the

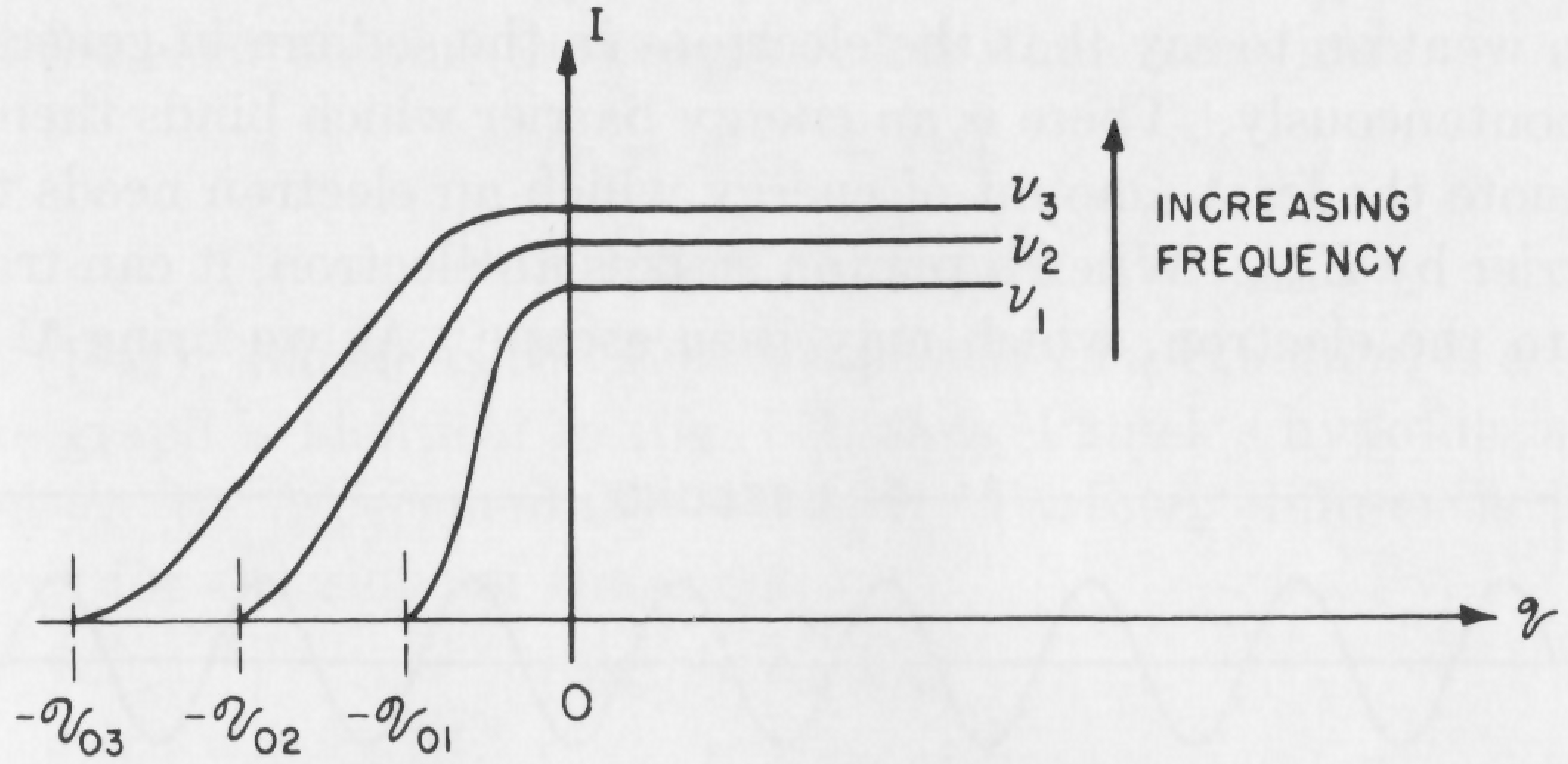


FIG. 1-8. The effect of increasing frequency on the current vs. voltage curves for the arrangement of Fig. 1-6. (From Sproull, *op. cit.*)

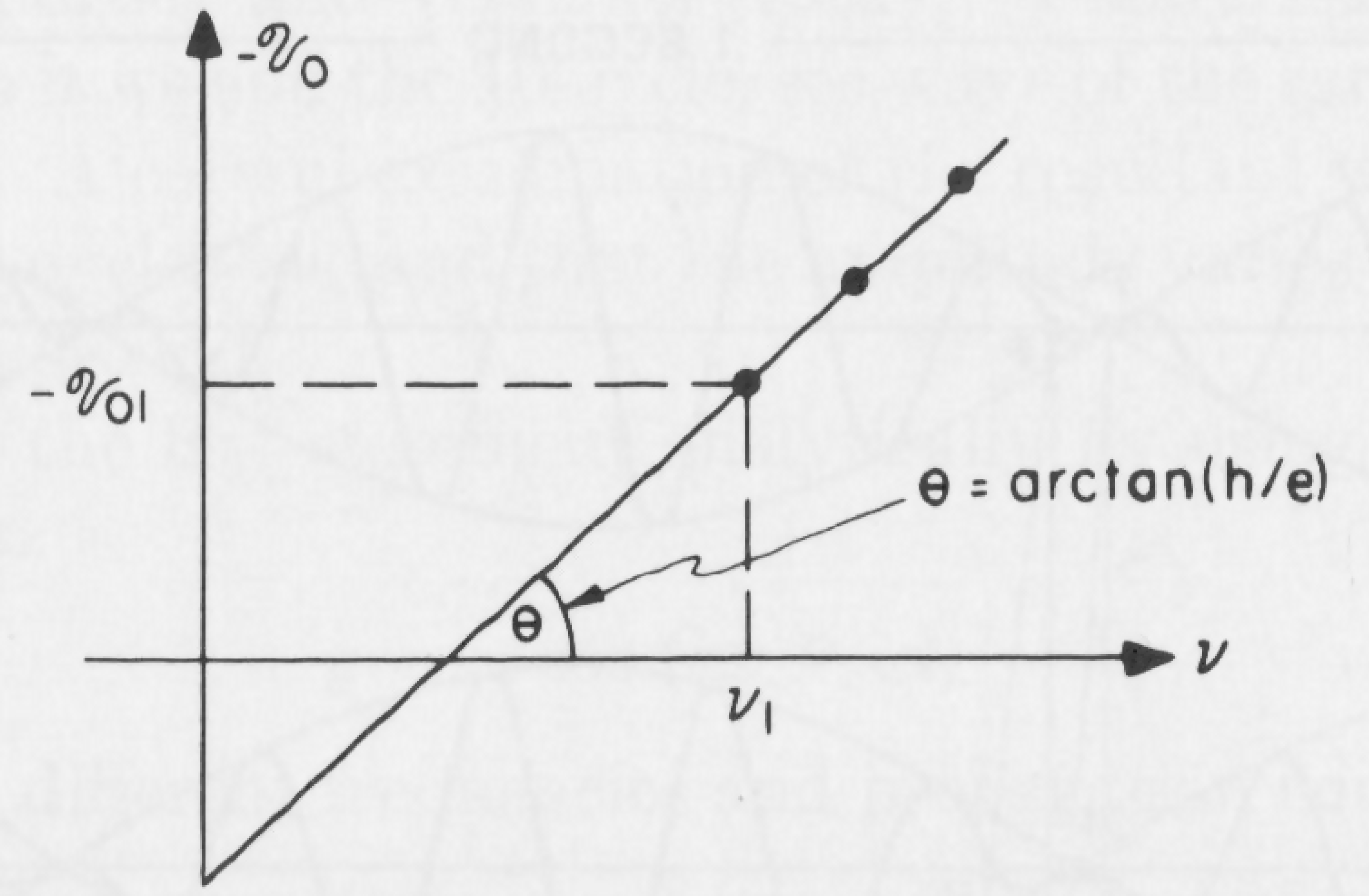


FIG. 1-9. Cut-off voltage as a function of frequency, as determined from Fig. 1-8. (From Sproull, *op. cit.*)

straight line of Fig. 1-9, where we have indicated how the points represent the pairs of variables $(\nu_1, -\mathcal{V}_{01})$, etc. Einstein explained these results with a remarkable assumption of Planck called the *quantum theory*, which states that

(1) Energy is not indefinitely divisible; the smallest unit of energy which may be transferred in a physical process is called a *quantum*. In particular, light is a form of energy and a quantum of light is called a *photon*.

(2) The energy E of a photon is proportional to the frequency ν of the light, or

$$E = h\nu \tag{1-40}$$

where h is called *Planck's constant* and has the value

$$h = 6.6 \times 10^{-34} \text{ joule-sec}$$

Einstein went on to say that the electrons in the sodium in general do not leave spontaneously. There is an energy barrier which binds them and we shall denote the least amount of energy which an electron needs to escape this barrier by $E_{\min}$. When a photon strikes an electron, it can transfer its energy to the electron, which may then escape. As we bring $\mathcal{V}$ down to

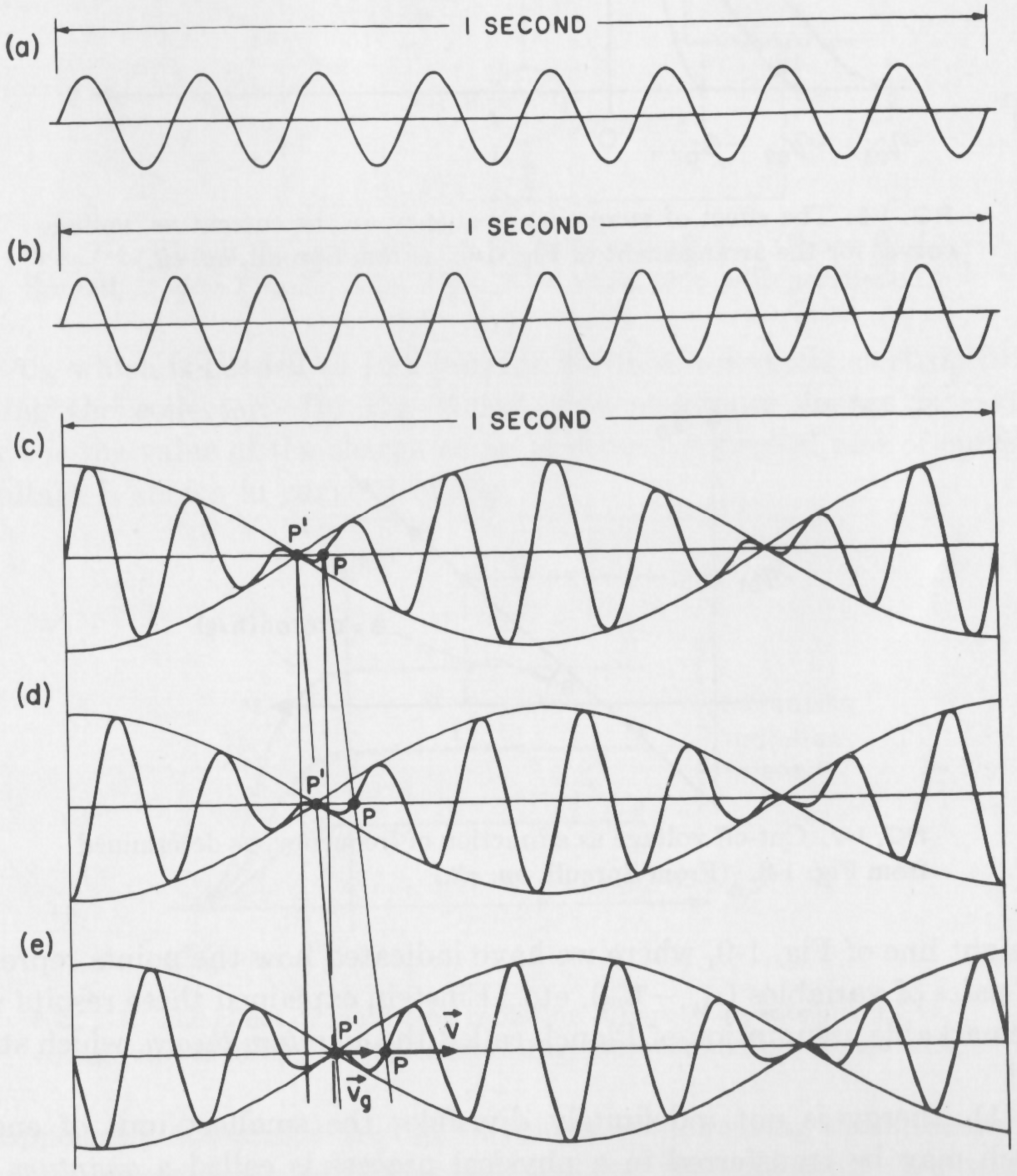


FIG. 1-10. (a) An eight-cycle/sec sine wave, and (b) a ten-cycle/sec wave with the same amplitude. (c) The sum of the two waves. (d) The sum after one-quarter of a cycle, and (e) the sum after one-half of a cycle. (From G. Shortley and D. Williams, *Elements of Physics*, 3rd ed., Prentice-Hall, Englewood Cliffs, N.J., 1961, and J. D. Kraus, *Electromagnetics*, McGraw-Hill, New York, 1953, with permission.)

$-\mathcal{V}_0$, slowly reducing the current to zero, the most energetic electrons — the ones that require only $E_{\min}$ to escape — are the only ones to reach the collector. Equating the energy of the photon to the energy which it takes to release the electron plus the energy it has when free yields

$$h\nu = E_{\min} + e\mathcal{V}_0$$

or

$$\mathcal{V}_0 = (h\nu/e) - (E_{\min}/e) \qquad (1\text{-}41)$$

Equation (1-41), known as the *Einstein photoelectric equation*, is a straight line whose graph is identical to Fig. 1-9; thus, Planck's hypothesis can be considered to be experimentally verified. Further, differentiating Eq. (1-41) gives for the slope of the graph

$$d\mathcal{V}_0/d\nu = h/e$$

and this value for the slope experimentally determines the value of Planck's constant h given above.

The remaining basic ideas we shall cover in this section involve a deeper understanding of wave motion. Let us extend the discussion of the previous section by considering the interaction of two waves. We shall use an example by Shortley and Williams (1961).[1] Figure 1-10a shows an 8-cycle/sec wave; to it we add the 10-cycle/sec wave of the same amplitude to obtain Fig. 1-10c. A careful examination of the resultant wave shows that the frequency is 9 cycles/sec and that the amplitude varies from zero to its maximum twice a second.

We can prove the last statement analytically by using the solution to the wave equation

$$y = a \cos (kx - \omega t) \qquad (1\text{-}38)$$

For waves of two different frequencies and propagation constants

$$y = a[\cos (k_1 x - \omega_1 t) + \cos (k_2 x - \omega_2 t)] \qquad (1\text{-}42)$$

where $k_1 = 2\pi/\lambda_1$ and $k_2 = 2\pi/\lambda_2$. It is convenient to express these quantities in terms of their averages by letting

$$\omega_1 = \omega_0 - \Delta\omega \qquad \omega_2 = \omega_0 + \Delta\omega$$

where

$$\omega_0 = (\omega_1 + \omega_2)/2 \qquad \Delta\omega = (\omega_2 - \omega_1)/2$$

Using these values, Eq. (1-42) becomes

$$y = [a \cos (\Delta k x - \Delta\omega t)] \cos (k_0 x - \omega_0 t)$$

where the term in square brackets is a variable amplitude in place of the constant amplitude of Eq. (1-38). As shown in Fig. 1-10c, the *magnitude* of the amplitude has a maximum twice each second. If the ear responded to frequencies as low as these, an observer would hear a 9-cycle/sec note with two *beats* per second.

Figure 1-10d, taken from Kraus (1953),[2] shows the same wave a quarter of a cycle later; Fig. 1-10e shows it after a half-cycle. The motion of the 9-cycle/sec wave is indicated by the successive positions of P, while P' shows the motion of the beats or the *envelope*. The velocity v of P, which represents the motion of an individual wave, is called the *phase velocity*. The velocity v_g of P' represents the motion of the group of waves, hence the name *group velocity*, and v_g is smaller than v. Kraus points out that the difference between phase and group velocities is nicely illustrated by the motion of a caterpillar. The humps on his back go forward with phase velocity, which may be quite rapid, but his motion over the ground is at the group velocity.

An expression for the group velocity comes by considering the value of $(\Delta kx - \Delta\omega t)$ for a point like P', which may be

$$\Delta kx - \Delta\omega t = \pi/2$$

If we differentiate both sides with respect to t, the velocity of P' will then be

$$v_g = dx/dt = \Delta\omega/\Delta k$$

or

$$1/v_g = dk/d\omega$$

in the limit. This may also be written

$$v_g = \frac{d\omega}{dk} = \frac{d\,(2\pi\nu)}{d\left(\dfrac{2\pi}{\lambda}\right)} = -\lambda^2\,\frac{d\nu}{d\lambda} \tag{1-43}$$

Let us return to Planck's quantum theory, expressed by the equation

$$E = h\nu \tag{1-40}$$

and add an additional thought to the idea that light energy exists in indivisible units or photons. Following a development of Pearson (1961),[3] we regard the photon as a particle or corpuscle of energy and associate a mass with it. The total energy is the sum of the kinetic and potential energies, or

$$E = h\nu = \tfrac{1}{2}\,mv^2 + V \tag{1-44}$$

If we postulate that the group velocity of the photon equals its particle velocity, then

$$-\lambda^2\,\frac{d\nu}{d\lambda} = v$$

Combining this with Eq. (1-44) yields

$$\frac{d\nu}{d\lambda} = \frac{1}{h}\,mv\,\frac{dv}{d\lambda} = \frac{-v}{\lambda^2}$$

Integrating and setting the arbitrary constant equal to zero gives

$$\lambda = h/mv = h/p \tag{1-45}$$

This result is known as *de Broglie's hypothesis* and is the statement of the idea of wave-particle duality; it connects the wave length of a photon regarded as a wave with the mass of the photon regarded as a particle. The same concept has been extended to the electron, which is also considered to be a dual entity, and Eq. (1-45) has been verified experimentally for electrons by Davisson and Germer.

These results are also related to relativity theory. The velocity of a photon is connected to the wave length and frequency by $c = \lambda\nu$. This, combined with Eqs. (1-44) and (1-45), gives

$$E = mc^2$$

which is the *Einstein equation for mass-energy equivalence*. Applying this to an electron wave yields

$$\lambda = \frac{h}{mv} = \frac{(mc^2/\nu)}{mv}$$

or

$$(v)(\nu\lambda) = (v)(\omega/k) = c^2 \tag{1-46}$$

This equation says that the product of the particle velocity and the phase velocity of the associated wave is a constant. Putting Eq. (1-46) into Eq. (1-43) and equating the group velocity and particle velocity gives

$$v_g = -\lambda^2\,\frac{d}{d\lambda}\left(\frac{c^2}{v\lambda}\right) = \frac{c^2}{v} - \lambda\,\frac{d}{d\lambda}\left(\frac{c^2}{v}\right) = v$$

Integrating yields

$$\frac{1}{\lambda} = \frac{Cv}{(1 - v^2/c^2)^{1/2}}$$

where C is a constant of integration. Hence

$$E = hc^2/\lambda v = hc^2C/(1 - v^2/c^2)^{1/2} = mc^2$$

Then

$$m = \frac{m_0}{(1 - v^2/c^2)^{1/2}} \tag{1-47}$$

where $m_0 = hC$ is called the *rest-mass*. Equation (1-47) shows that mass is a function of velocity.

PROBLEM 1-8 Find m when $v = 0$, $0.1c$, $0.9c$, c, and $1.1c$. Interpret the results.

The definition of force Eq. (1-4), shows that when mass depends on velocity

$$F = dp/dt = d(mv)/dt = m\,dv/dt + v\,dm/dt$$

or

$$F = \frac{m_0}{\sqrt{1 - v^2/c^2}}\,\frac{dv}{dt} + \frac{v^2 m_0/c^2}{(1 - v^2/c^2)^{3/2}}\,\frac{dv}{dt}$$

$$= \frac{m_0}{(1 - v^2/c^2)^{3/2}}\,\frac{dv}{dt}$$

The kinetic energy may be obtained, as in Section 1-3, by integrating

$$T = \int F\,dx = m_0 \int_0^v \frac{v\,dv}{(1 - v^2/c^2)^{3/2}} = m_0c^2\left(\frac{1}{\sqrt{1 - v^2/c^2}} - 1\right)$$

$$= mc^2 - m_0c^2$$

This may be written as

$$E = T + m_0c^2$$

where m_0c^2 is called the *rest-energy*. Using the binomial theorem, we also see that

$$T = m_0c^2(1 + v^2/2c^2 + \ldots) - m_0c^2$$

$$= \tfrac{1}{2}mv^2$$

provided $v \ll c$, so the usual expression for kinetic energy actually represents an approximation valid at velocities much less than the speed of light.

Another form of Eq. (1-45) which is in common use comes by introducing the propagation constant $k = 2\pi/\lambda$ of Eq. (1-34) and the special symbol $\hbar = h/2\pi$ (read as "$\hbar$-bar"). Then

$$p = 2\pi\hbar/\lambda = \hbar k \tag{1-48}$$

Equation (1-40) becomes

$$E = h\nu = h\omega/2\pi = \hbar\omega \tag{1-49}$$

1-7 The Schroedinger Equation

Having described the electron as something which shows both wave and particle properties, we are now ready to consider the problem of expressing its motion by an equation. In Section 1-5, we showed that a single vibrating particle of mass m obeys

$$\frac{d^2x}{dt^2} + \omega^2 x = 0 \tag{1-28}$$

and the wave created by a group of vibrating particles is described by

$$\frac{\partial^2 y}{\partial t^2} = v^2\,\frac{\partial^2 y}{\partial x^2} \tag{1-39}$$

We would like an analogous equation which covers the motion of a dual entity, such as an electron. To achieve this, we will first replace the cosine solution

$$y = a\cos(kx - \omega t) \tag{1-38}$$

of Eq. (1-39) with the equivalent exponential solution

$$y = a\,e^{i(kx - \omega t)} \tag{1-50}$$

PROBLEM 1-9 Verify that Eq. (1-50) is a solution of Eq. (1-39).

Now we substitute Eqs. (1-48) and (1-49) into Eq. (1-50), changing it to

$$y = a\,e^{i(xp - Et)/\hbar} = \Psi \tag{1-51}$$

where we replace y by Ψ to emphasize that Eq. (1-51), which incorporates Planck's and de Broglie's hypotheses, refers to a wave-particle motion, rather than to a pure wave.

PROBLEM 1-10

(a) Calculate $\partial^2\Psi/\partial x^2$ and $\partial\Psi/\partial t$.

(b) Make use of the fact that the total energy E of the electron is the sum of its kinetic energy T and its potential energy V

$$E = T + V \tag{1-52}$$

and combine this with (a) and with Eqs. (1-3) and (1-8) to show that

$$\frac{-\hbar^2}{2m}\,\frac{\partial^2\Psi}{\partial x^2} + V\Psi - i\hbar\,\frac{\partial\Psi}{\partial t} = 0 \tag{1-53}$$

This is the *time-dependent Schroedinger*, or *wave, equation*, and it is the differential equation which governs the motion of an electron. We would like to re-emphasize here the statement made in Section 1-1: the steps which led to Eq. (1-53) are not to be construed as a rigorous derivation, but merely as an attempt to make this equation plausible.

Since Eq. (1-53) is a partial differential equation involving x and t as independent variables, we will attempt to simplify it by using the standard procedure known as the method of separation of variables. Assume that Ψ

can be expressed as the product of two new unknown functions: $\psi(x)$, which depends only on x, and $T(t)$, which depends only on t. Then

$$\Psi(x, t) = \psi(x)\,T(t) \tag{1-54}$$

Since $\psi(x)$ is not a function of t,

$$\frac{\partial \Psi}{\partial t} = \psi \frac{dT}{dt}$$

Similarly,

$$\frac{\partial^2 \Psi}{\partial x^2} = T \frac{d^2\psi}{dx^2}$$

Using these, Eq. (1-53) becomes, after rearrangement,

$$\frac{-\hbar^2}{2m\psi}\frac{d^2\psi}{dx^2} + V = \frac{i\hbar}{T}\frac{dT}{dt} \tag{1-55}$$

This is an equation in which one side depends only on x and the other side depends only on t. The only way in which this is possible is for both sides to be equal to a constant, C, giving two ordinary differential equations, of which one is

$$i\hbar\, dT/dt = CT$$

and this can be integrated to give

$$T = e^{Ct/i\hbar}$$

so that

$$\Psi = \psi(x)e^{-iCt/\hbar}$$

where we have used $i = -1/i$.

If this Ψ is the same as the Ψ of Eq. (1-51), we must have

$$E = C$$

This makes the other equation taken from Eq. (1-55)

$$\frac{-\hbar^2}{2m}\frac{d^2\psi}{dx^2} + V\psi = E\psi \tag{1-56}$$

This is called the *Schroedinger amplitude equation*, or simply the Schroedinger equation. It should be emphasized that the quantity ψ, called a *wave function*, does not have a simple physical interpretation, whereas the corresponding classical quantity, y, of Eqs. (1-38) or (1-50) represents the height of the wave at a given place and time.

It should be realized that the Schroedinger equation is really many differential equations, since the potential energy V can take a number of forms. It can be a constant, including zero; it can arise from a harmonic or an inverse square force; it can even be a complicated periodic function in a real crystal. In the sections that follow, we shall show how Eq. (1-56) can be solved for some of these possible forms of V.

1-8 The Quantum Oscillator

The vibrating mass m on the end of a spring of Hooke's law constant κ which we considered in Section 1-5 is called a classical harmonic oscillator. If a particle of atomic size moves under the influence of a harmonic force, its motion is governed by the Schroedinger equation [Eq. (1-55)], and it is called a quantum oscillator. A harmonic force was defined by Eq. (1-27)

$$F = -\kappa x \tag{1-27}$$

and the corresponding potential energy is given by Eq. (1-5)

$$V = -\int F\,dx = \tfrac{1}{2}\kappa x^2 \tag{1-57}$$

Hence, Eq. (1-56) becomes

$$\frac{d^2\psi}{dx^2} + \frac{2m}{\hbar^2}\left(E - \frac{\kappa x^2}{2}\right)\psi = 0 \tag{1-58}$$

If we let $\alpha = 2mE/\hbar^2$ and $\beta^2 = m\kappa/\hbar^2$, then Eq. (1-58) can be written

$$\frac{d^2\psi}{dx^2} + (\alpha - \beta^2x^2)\psi = 0 \tag{1-59}$$

A solution of Eq. (1-58) is

$$\psi = e^{-\beta x^2/2}$$

provided $\alpha = \beta$.

PROBLEM 1-11 Verify the result given above.

This leads us to try a solution

$$\psi(x) = e^{-\beta x^2/2} f(x) \tag{1-60}$$

where $f(x)$ is a new function to be determined. Substituting Eq. (1-60) into Eq. (1-59) leads to

$$\frac{d^2f}{dx^2} - 2\beta x\frac{df}{dx} + (\alpha - \beta)f = 0 \tag{1-61}$$

This is known as the *Hermite equation* (after a 19th century mathematician and astronomer), and it has an infinite number of solutions, as follows

$$\begin{array}{ll} f = 1 & \text{if} \quad \alpha = \beta \\ f = x & \text{if} \quad \alpha = 3\beta \\ f = 2\beta x^2 - 1 & \text{if} \quad \alpha = 5\beta \\ \quad\vdots & \qquad\vdots \end{array} \tag{1-62}$$

From the definitions of α and β, we have for $\alpha = \beta$

$$E = \frac{\sqrt{\kappa/m}}{2\hbar} = \frac{\hbar\omega}{2} = \frac{h\nu}{2}$$

and for $\alpha = 3\beta$, $\alpha = 5\beta$, etc.,

$$E = 3h\nu/2, \qquad E = 5h\nu/2, \qquad \text{etc.}$$

or, in general,

$$E_n = (n + \tfrac{1}{2})h\nu, \quad n = 0, 1, 2, \ldots \tag{1-63}$$

where the integer n is called a *quantum number*. The Schroedinger equation has thus shown us that the harmonic oscillator is allowed only certain energies, known as *energy levels*, and the energy of this system is said to be *quantized*. Note that the levels are 1 quantum, $h\nu$, apart.

Now let us consider the wave functions, which, by Eqs. (1-60) and (1-62), are

$$\begin{aligned} \psi_0 &= e^{-\beta x^2/2} \\ \psi_1 &= xe^{-\beta x^2/2} \\ \psi_2 &= (2\beta x^2 - 1)\, e^{-\beta x^2/2} \\ &\;\;\vdots \end{aligned} \tag{1-64}$$

From Eq. (1-64), we expect that the functions ψ_3, ψ_4, etc. will involve more complicated polynomials in x, which turns out to be the case. For reasons which will soon be apparent, we show a plot of the square of the function ψ_{10} in Fig. 1-11. In this figure, the solid line is the curve ψ_{10}^2 vs. x. The dotted line is the average value of ψ_{10}^2, obtained by taking the area under each "hump" and dividing by the corresponding length of the base line. The points $\pm\, a$ on the x axis represent the amplitude of a classical oscillator having the same total energy. Comparing this curve with Fig. 1.4, we see that there is a close resemblance between the probability curve for the classical vibrating mass and for the average value of ψ_{10}^2 for the particle which obeys the Schroedinger equation. It seems, therefore, very reasonable to postulate that $\psi_n^2(x)$, where $n = 0, 1, 2, \ldots$, is the *probability per unit length* of finding an oscillator of energy $E_n = (n + \frac{1}{2})h\nu$ in a small length dx located at x. This quantity is also called the *probability density*. It should be noted that it is the average of ψ_n^2 which resembles Fig. 1-4, and not ψ_n^2 itself, for we see that ψ_n^2 is not zero at $x = \pm\, a$, and hence there is a finite probability of finding the quantum particle in places which are not accessible to the classical one. We shall find this idea recurring further on.

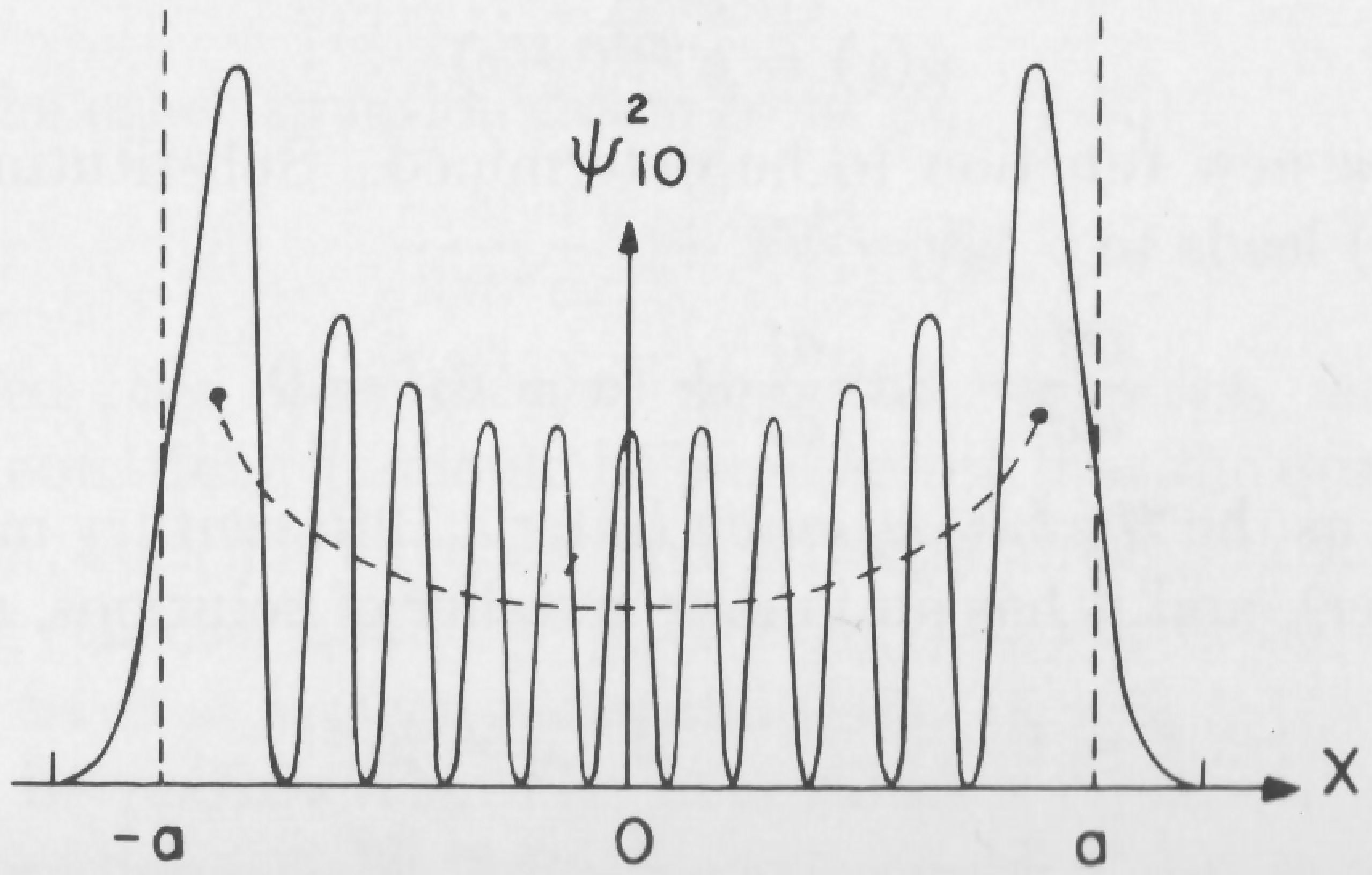


FIG. 1-11. A plot of ψ_{10}^2 as a function of x. The dotted curve represents the average value of ψ_{10}^2 and the points $x = \pm a$ are the maximum displacement of a classical oscillator with the same mass and spring constant. (From L. Pauling and E. B. Wilson, Jr., *Introduction to Quantum Mechanics*, McGraw-Hill, New York, 1935, with permission.)

PROBLEM 1-12

(a) Prove that the total energy E and the amplitude a of an oscillator are related by

$$a = \sqrt{2E/k}$$

(b) Use this result to show that the points $x = \pm\, a$ of Fig. 1-11 are located a distance $(21\, h/4\pi^2 m\nu)^{1/2}$ from the origin.

1-9 Potential Wells

The next type of a potential energy form which we wish to consider is

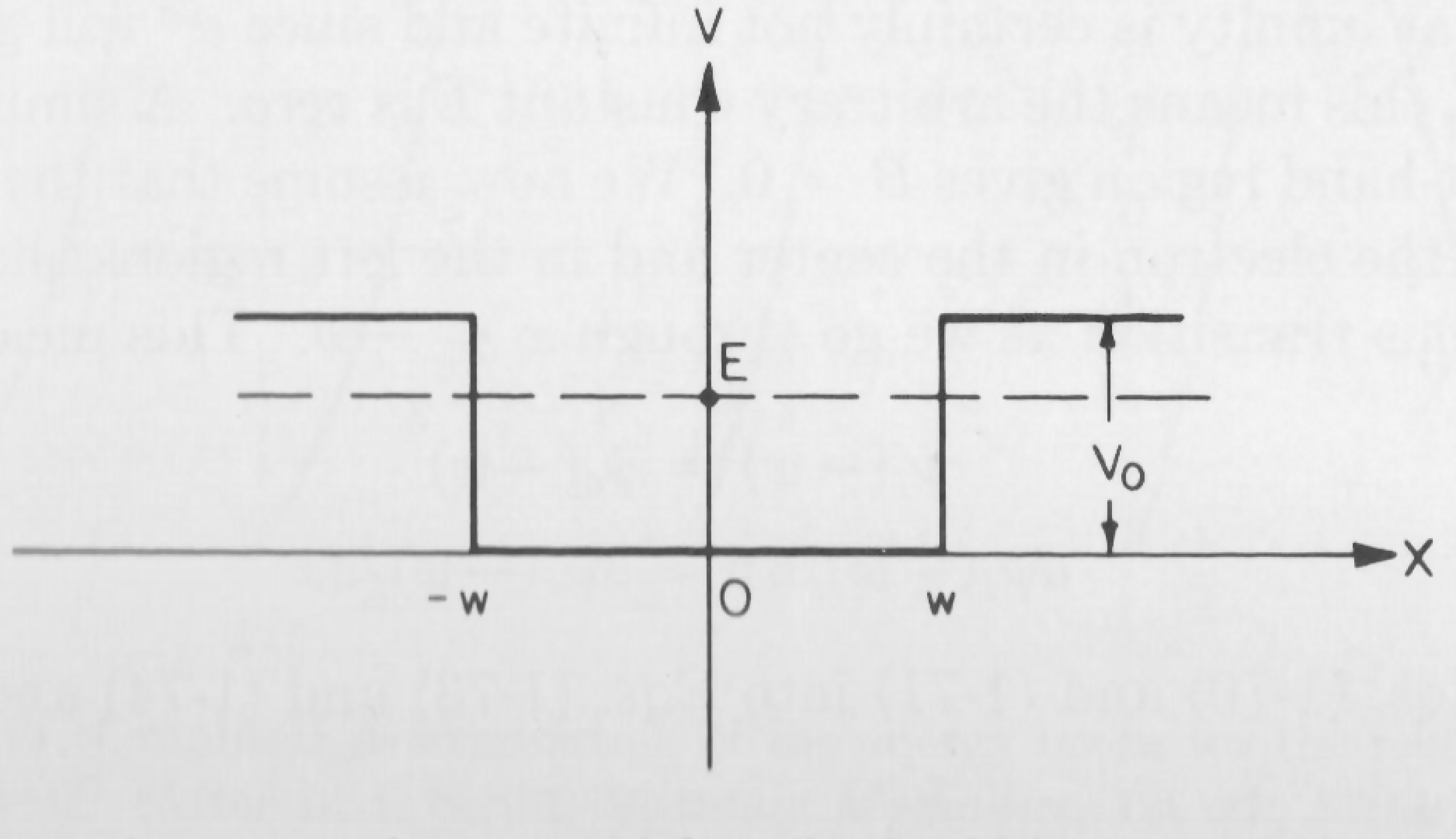


FIG. 1-12. A rectangular potential well of width $2w$ and depth V_0.

called a *rectangular potential well;* it is shown in Fig. 1-12. This dependence of V on x is

$$V = \begin{cases} V_0 \text{ for } & x \leqq -w \\ 0 \text{ for } & -w \leqq x \leqq w \\ V_0 \text{ for } & x \geqq w \end{cases} \tag{1-65}$$

Let us assume that the total energy E of the particle is less than the height V_0 of the well, so that the particle would not be expected to have enough energy to escape from the well. We shall soon see that, in spite of this restriction, there will be a finite probability of finding the particle outside the well, just as in the previous section we found it beyond the classical limits. We now divide the interval $-\infty < x < \infty$ into three regions: left, central, and right. There will be an equation of the form of Eq. (1-56) for each region.

$$d^2\psi_l/dx^2 - \beta^2\psi_l = 0 \quad (x \leqq -w) \tag{1-66}$$

$$d^2\psi_c/dx^2 + \alpha^2\psi_c = 0 \quad (-w \leqq x \leqq w) \tag{1-67}$$

$$d^2\psi_r/dx^2 - \beta^2\psi_r = 0 \quad (x \geqq w) \tag{1-68}$$

where

$$\beta^2 = 2m(V_0 - E)/\hbar^2, \qquad \alpha^2 = 2mE/\hbar^2 \tag{1-69}$$

and both α^2 and β^2 are positive.

The solutions to Eqs. (1-66), (1-67), and (1-68) are, respectively,

$$\psi_l = Ae^{\beta x} + Be^{-\beta x} \tag{1-70}$$

$$\psi_c = Ce^{i\alpha x} + De^{-i\alpha x} \tag{1-71}$$

$$\psi_r = Ee^{\beta x} + Fe^{-\beta x} \tag{1-72}$$

where A through F are arbitrary constants. Since the probability of finding a particle at infinity is certainly not infinite and since $e^{\beta x}$ will get very large as $x \to \infty$, this means the arbitrary constant E is zero. A similar argument for the left-hand region gives $B = 0$. We now assume that the probabilities of finding the electron in the center and in the left regions should undergo a continuous transition as we go through $x = -w$. This means

$$\psi_l(-w) = \psi_c(-w) \tag{1-73}$$

and

$$d\psi_l(-w)/dx = d\psi_c(-w)/dx \tag{1-74}$$

Putting Eqs. (1-70) and (1-71) into Eqs. (1-73) and (1-74) and eliminating A gives

$$D/C = -(\beta - i\alpha)\, e^{-2i\alpha w}/(\beta + i\alpha) \tag{1-75}$$

Library
State College
Indiana, Penna.
537.62 N944
c.1

PROBLEM 1-13

(a) Verify Eq. (1-75).

(b) Find the corresponding equation at $x = w$.

(c) Prove $D/C = \pm 1$.

(d) From (c), show that

$$\beta w = -\alpha w \cot \alpha w \tag{1-76}$$

$$\beta w = \alpha w \tan \alpha w \tag{1-77}$$

using the identity from Appendix A1

$$e^{\pm ix} = \cos x \pm i \sin x$$

Equations (1-76) and (1-77) cannot be solved analytically for the unknowns

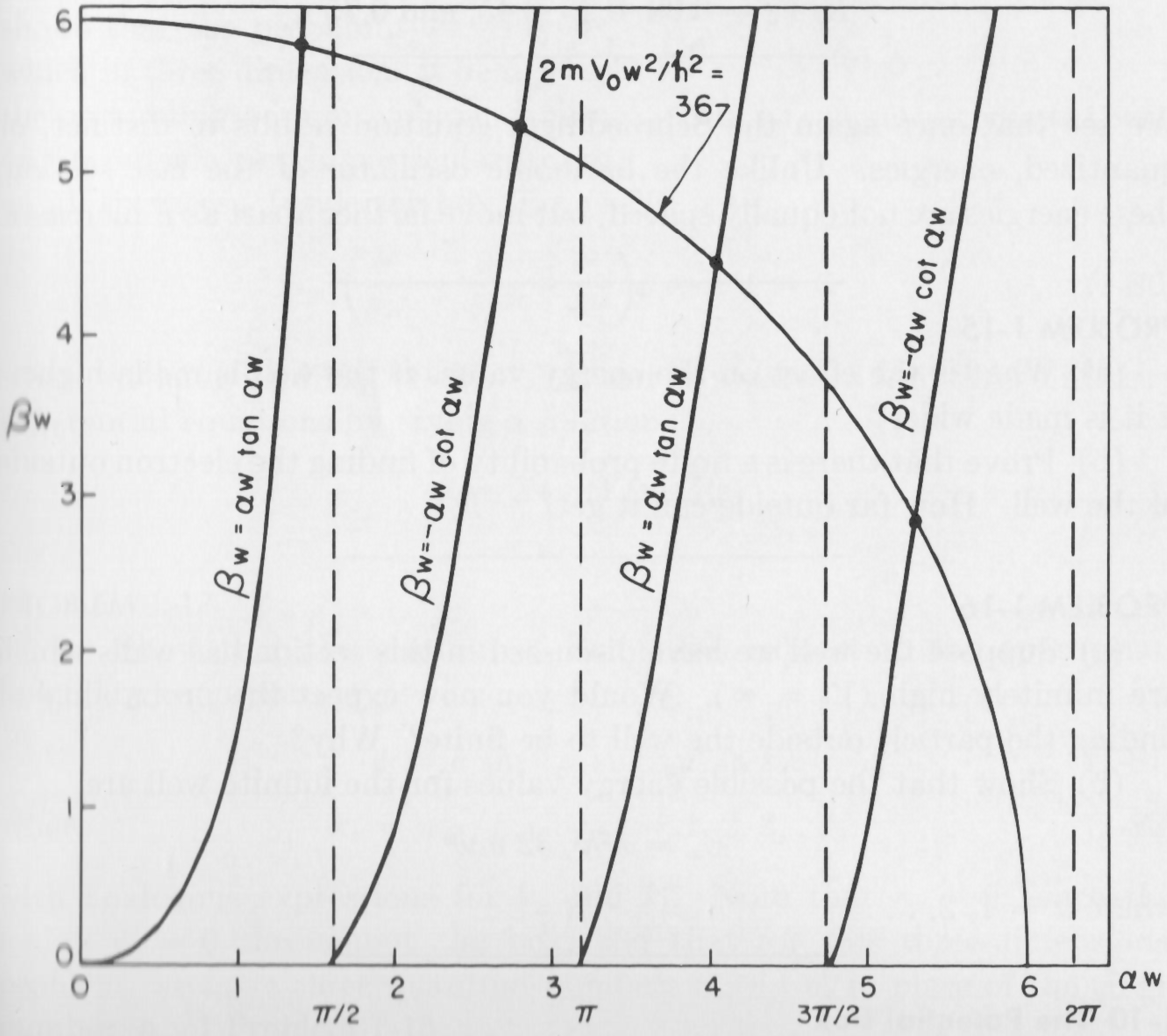


FIG. 1-13. Graphical determination of the energy levels for the rectangular well. (After L. I. Schiff, *Quantum Mechanics*, 1st ed., McGraw-Hill, New York, 1949, with permission.)

β and α. However, they can be graphically solved by combining the two equations (1-69) to get the equation of a circle.

$$\alpha^2 w^2 + \beta^2 w^2 = 2mV_0 w^2/\hbar^2 \tag{1-78}$$

The intersections of this circle with the curves of Eqs. (1-76) and (1-77) are the only values of α and β which satisfy both the definitions of equations (1-69) and the boundary conditions. Figure 1-13 shows the appearance of these curves using the numerical example of Rojansky (1938)[4] for which $2mV_0 w^2/\hbar^2 = 36$. Having found αw and βw, we can then determine the corresponding values of E.

PROBLEM 1-14

(a) Find a formula for E in terms of V_0, αw, and βw.

(b) For the example of Fig. 1-13, show that the approximate values of the energy levels are

$$E/V_0 = 0.05, 0.20, 0.45, \text{and } 0.75$$

We see that once again the Schroedinger equation results in distinct, or quantized, energies. Unlike the harmonic oscillator of the last section, these energies are not equally spaced, but move farther apart as E increases.

PROBLEM 1-15

(a) What is the effect on the energy values if the well is made higher; if it is made wider?

(b) Prove that there is a finite probability of finding the electron outside of the well. How far outside can it get?

PROBLEM 1-16

(a) Suppose the well we have discussed in this section has walls which are infinitely high ($V_0 = \infty$). Would you now expect the probability of finding the particle outside the well to be finite? Why?

(b) Show that the possible energy values for the infinite well are

$$E_n = n^2h^2/32\,mw^2$$

where $n = 1, 2, \ldots$

1-10 The Potential Box

Equation (1-56) governs the motion of an electron confined to the x-axis. If the electron is allowed to move in an arbitrary direction, we would

expect its motion to depend on y and z in the same manner as it depends on x, and the three-dimensional Schroedinger equation should therefore be

$$\frac{-\hbar^2}{2m}\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}\right)\psi + V\psi = E\psi$$

or

$$\frac{-\hbar^2}{2m}\nabla^2\psi + V\psi = E\psi \tag{1-79}$$

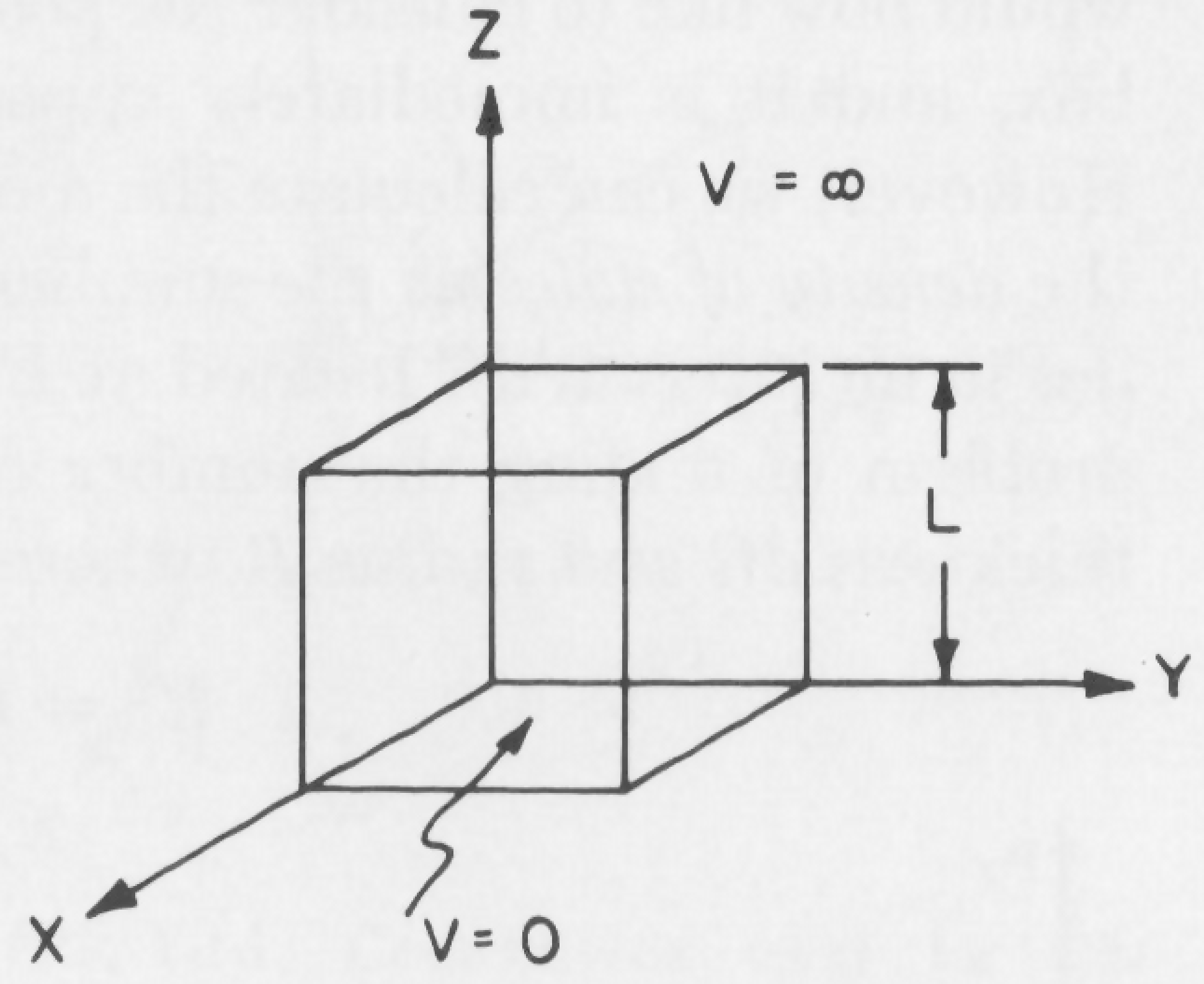


FIG. 1-14. A cubical potential box.

where the symbol ∇^2 is defined in Appendix A3. Let us solve this equation for a cubical *potential box*, defined as a region whose potential energy is zero inside the cube of side L and infinite everywhere else (*see* Fig. 1-14). Since it would take an infinite amount of work to get an electron outside the box, the reasoning used in Problem 1-16 shows that the probability-density ψ^2, which in three dimensions is defined as the probability per unit volume, is zero on the outside, and in particular is zero on all six faces. Inside the box, Eq. (1-79) is

$$\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}\right)\psi + k^2\psi = 0 \tag{1-80}$$

where we have used Eq. (1-48). This can be separated into three ordinary differential equations by trying a solution

$$\psi = X(x)\,Y(y)\,Z(z)$$

PROBLEM 1-17

(a) Show that the solution of Eq. (1-80) subject to the appropriate boundary conditions is

$$\psi = a \sin k_x x \sin k_y y \sin k_z z \tag{1-81}$$

where

$$k_x = n_x\pi/L, \qquad n_x = 1, 2, 3, \ldots \tag{1-82}$$

with analogous expressions for k_y and k_z. Note that $n_x \neq 0$, since this makes $\psi^2 = 0$ throughout the box, and that for this three-dimensional problem, we have three quantum numbers n_x, n_y, n_z in place of the single number, n, of Problem 1-16.

(b) Show that this solution is valid, provided

$$-(k_x^2 + k_y^2 + k_z^2) + k^2 = 0 \tag{1-83}$$

and that the energy levels are

$$E = \frac{\pi^2\hbar^2}{2mL^2}(n_x^2 + n_y^2 + n_z^2) \tag{1-84}$$

Each set of integers (n_x, n_y, n_z) in this problem determines both ψ and E, and these numbers are said to specify a possible *state* of the electron. We would now like to consider the possible states available to an electron in the box, and it is immediately apparent that there are an infinite number. However, we can calculate the number in a given range of energies. Define the *density of states* as the number dG of such possible states whose energy lies in an interval dE located at E. By Eq. (1-84), this is equivalent to the problem of finding the number of energy levels corresponding to a shell thickness dR and radius R, where

$$R^2 = n_x^2 + n_y^2 + n_z^2 \tag{1-85}$$

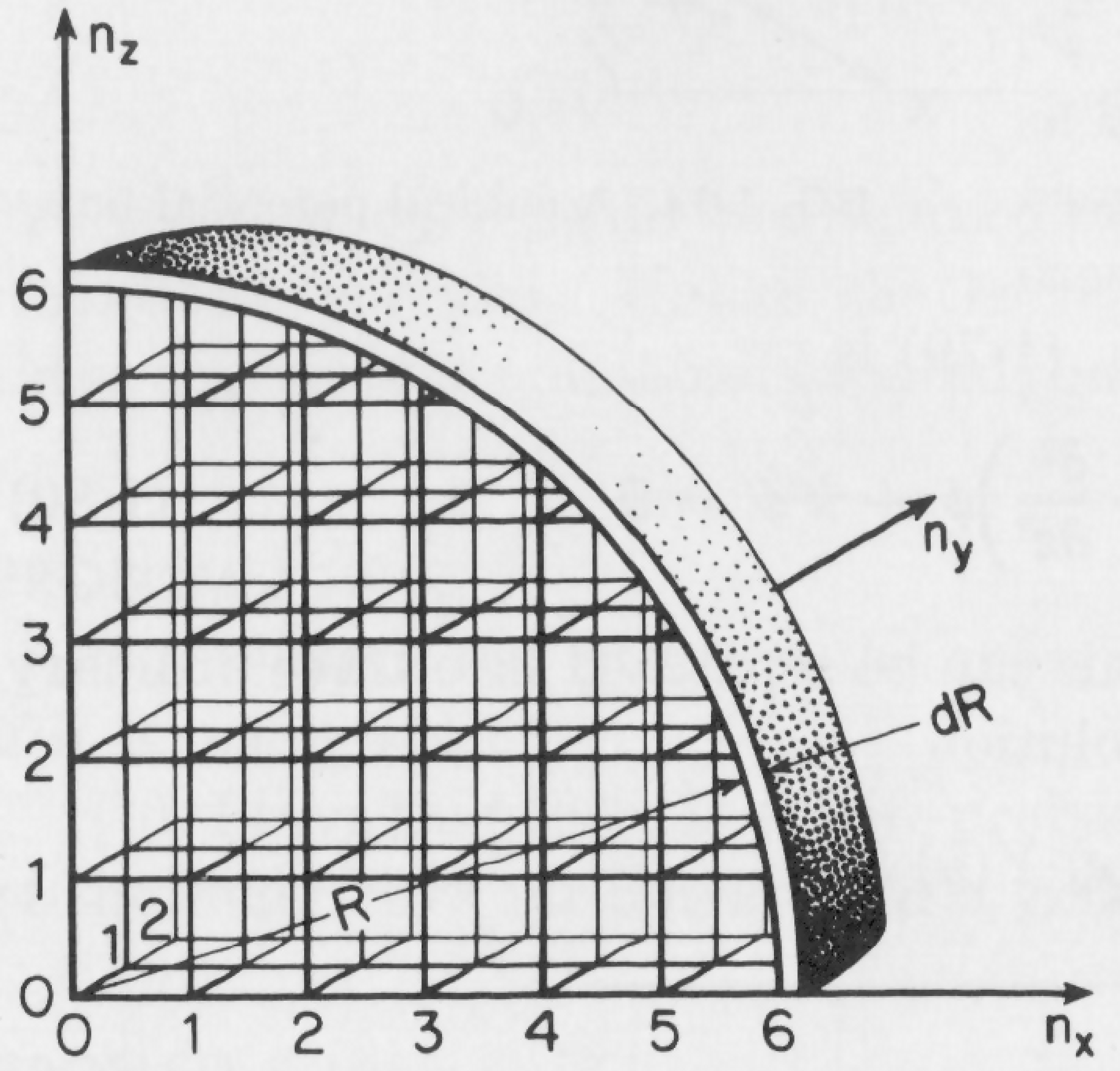


FIG. 1-15. Determination of the density of states corresponding to a thin shell of radius R and thickness dR.

In the n_x, n_y, n_z coordinate system of Fig. 1-15, there will be a value of E corresponding to every point whose coordinates are all integers. Since each point is equivalent to a volume of unity, and since n_x, n_y, and n_z are all positive, the number of points we are trying to calculate is identical to the volume of one octant of the shell of thickness dR, or

$$dG = (4\pi R^2\, dR/8) = (\pi R^2\, dR/2) \tag{1-86}$$

By Eq. (1-85),

$$R^2 = 2mL^2E/\pi^2\hbar^2 = p^2L^2/\pi^2\hbar^2$$

and $R\, dR = (p\, dp\, L^2/\pi^2\hbar^2)$

Hence, $$dG = (p^2L^3\, dp/2\pi^2\hbar^3) = (4\pi p^2V\, dp/h^3) \tag{1-87}$$

where V is the volume of the box.

PROBLEM 1-18 Express Eq. (1-87) in terms of energy rather than momentum by using $p^2 = 2mE$ to obtain

$$dG = 4\pi\, m^{3/2}\, (2E)^{1/2}V\, dE/h^3 \tag{1-88}$$

1-11 The Hydrogen Atom

We are now ready to consider a physical system for which the deductions based on the Schroedinger equation can be verified by experiment. This system is the *hydrogen atom*, which consists of a proton of charge e and an electron of charge $-e$ and mass m. We will regard the position of the proton as fixed at the origin, as shown in Fig. 1-16, and the electron as being at a point whose spherical coordinates are (r, θ, φ).* The force between the particles, by Eq. (1-12), is

$$F = -e^2/4\pi\epsilon_0 r^2$$

and the potential energy is then

$$V = -\int F\, dr = -e^2/4\pi\epsilon_0 r \tag{1-89}$$

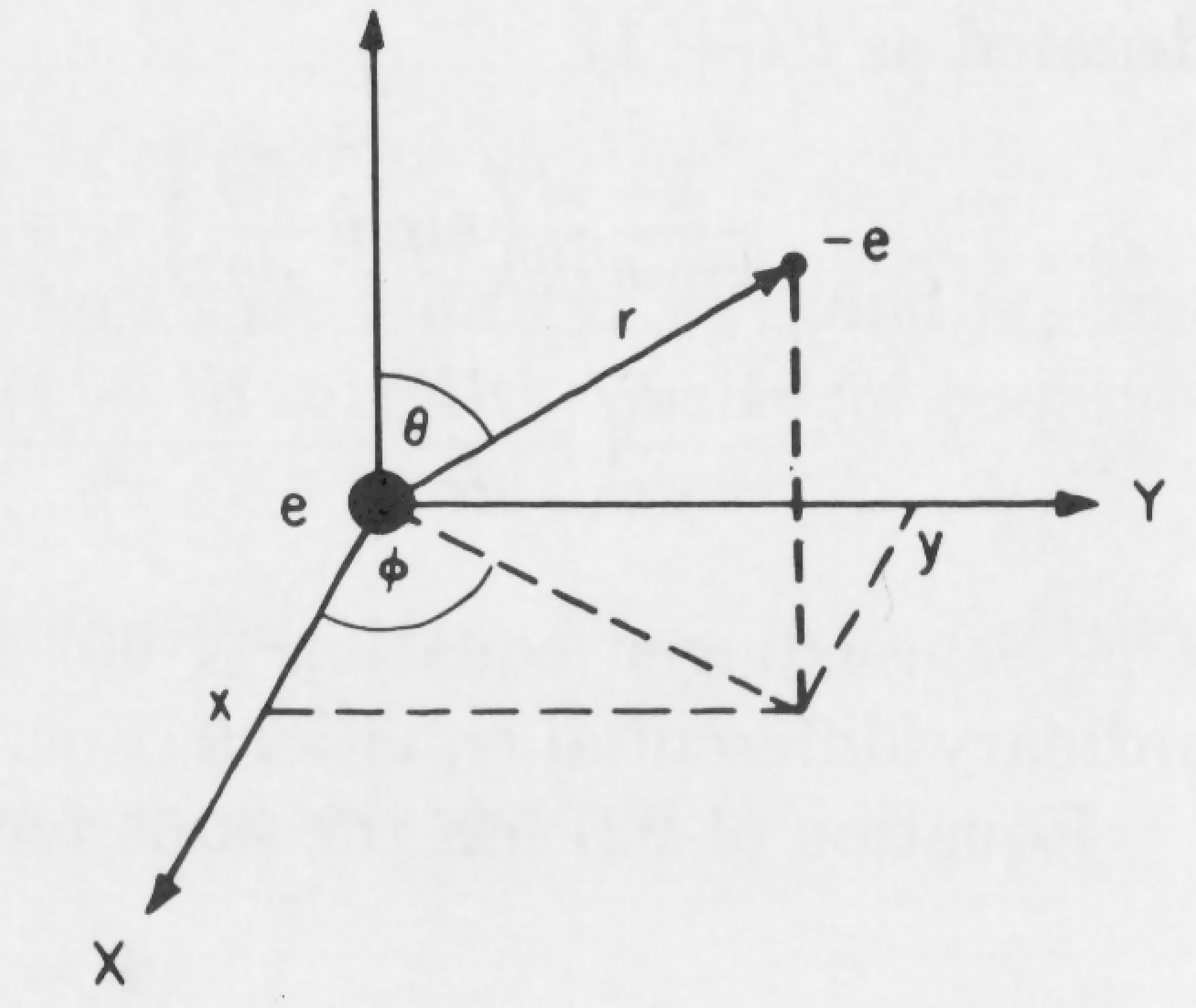


FIG. 1-16. Coordinates used for the hydrogen atom.

The Schroedinger equation [Eq. (1-79)] in spherical coordinates is

$$\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial\psi}{\partial r}\right) + \frac{1}{r^2\sin^2\theta}\frac{\partial^2\psi}{\partial\varphi^2} + \frac{1}{r^2\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial\psi}{\partial\theta}\right) + \frac{2m}{\hbar^2}[E - V(r)]\psi = 0 \tag{1-90}$$

This partial differential equation can be separated by letting

$$\psi(r, \theta, \varphi) = R(r)\Theta(\theta)\Phi(\varphi) \tag{1-91}$$

where $R(r)$ is a function of r alone, with similar definitions for $\Theta(\theta)$ and $\Phi(\varphi)$. Computing the appropriate derivatives of Eq. (1-91) and substituting into Eq. (1-90) gives

$$\frac{1}{r^2R}\frac{d}{dr}\left(r^2\frac{dR}{dr}\right) + \frac{1}{(r^2\sin^2\theta)\Phi}\frac{d^2\Phi}{d\varphi^2} + \frac{1}{(r^2\sin\theta)\Theta}\frac{d}{d\theta}\left(\sin\theta\frac{d\Theta}{d\theta}\right) + \frac{2m}{\hbar^2}[E - V(r)] = 0$$

By the use of a separation constant, which is written as m_l^2 for convenience, this can be broken up into the following two equations:

$$\frac{d^2\Phi}{d\varphi^2} = -m_l^2\Phi \tag{1-92}$$

---

*See the appendix for a description of the spherical coordinate system.

and
$$\frac{1}{R}\frac{d}{dr}\left(r^2\frac{dR}{dr}\right) - \frac{m_l^2}{\sin^2\theta} + \frac{1}{\sin\theta\,\Theta}\frac{d}{d\theta}\left(\sin\theta\frac{d\Theta}{d\theta}\right) + \frac{2m}{\hbar^2}[E - V(r)] = 0$$

The second of these can be further separated by the use of another constant, denoted as $l(l+1)$

$$\frac{1}{\sin\theta}\frac{d}{d\theta}\left(\sin\theta\frac{d\Theta}{d\theta}\right) - \frac{m_l^2}{\sin^2\theta}\,\Theta + l(l+1)\,\Theta = 0 \tag{1-93}$$

$$\frac{1}{r^2}\frac{d}{dr}\left(r^2\frac{dR}{dr}\right) - \frac{l(l+1)}{r^2}R + \frac{2m}{\hbar^2}[E - V(r)] = 0 \tag{1-94}$$

The Schroedinger equation (1-90) has thus been broken up into three ordinary differential equations [Eqs. (1-92), (1-93), and (1-94)].

Equation (1-92) has the same form as Eq. (1-28) and has solutions

$$\Phi = e^{\pm im_l\varphi} \tag{1-95}$$

where we have chosen the complex exponential form similar to that in Eq. (1-50).

PROBLEM 1-19 By the use of Appendix A1, show that the solution [Eq. (1-95)] is multivalued at $\varphi = 0$ or 2π unless m_l is an integer.

Hence, the quantum number m_l has the values

$$m_l = 0, \pm1, \pm2, \ldots$$

Also, in the case of complex functions, the definition ψ^2 for the probability density is modified to $\psi^*\psi = |\psi|^2$, where ψ^* is the complex conjugate of ψ, obtained by replacing i with $-i$.

Equation (1-93) is known as the *associated Legendre equation* and has physically meaningful solutions only when l is an integer and when $|m_l| \leq l$. These solutions are denoted by $\Theta_{l,m}(\theta)$ and the first few are

$$\left.\begin{array}{lll} & \Theta_{0,0} = 1 & \\ & \Theta_{1,0} = \cos\theta, \quad \Theta_{1,\pm1} = \sin\theta & \\ \Theta_{2,0} = (3\cos^2\theta - 1), & \Theta_{2,\pm1} = \sin\theta\cos\theta, & \Theta_{2,\pm2} = \sin^2\theta \end{array}\right\} \tag{1-96}$$

Equation (1-94) can be solved by substituting

$$R(r) = e^{-\alpha r}\,r^l\,L(r)$$

where $\alpha^2 = -2mE/\hbar^2$ and $L(r)$ is a new function to be determined. The equation which $L(r)$ obeys is known as the *associated Laguerre equation* and has physically meaningful solutions only if

$$n^2 = -me^4/8h^2E\,\epsilon_0^2 \tag{1-97}$$

where $n = 1, 2, 3, \ldots$

and if
$$l = 0, 1, 2, \ldots, n-1$$

This problem thus involves the three quantum numbers n, l, and m_l, but the energy levels depend only on n, which is called the *principal quantum number*. These values are then

$$E_n = -me^4/8n^2h^2\epsilon_0^2 \tag{1-98}$$

PROBLEM 1-20

(a) Show that Eq. (1-98) can be written

$$E_n = -(13.6/n^2)\,eV$$

(b) What measurable physical property of the atom does E_1 represent?

The first few associated Laguerre functions $L_{n,l}(r)$ are

$$\left.\begin{array}{lll} & L_{1,0} = 1 & \\ & L_{2,0} = (1 - \alpha r), \quad L_{2,1} = 1 & \\ L_{3,0} = (3 - 6\alpha r + 2\alpha^2r^2), & L_{3,1} = (2 - \alpha r), & L_{3,2} = 1 \end{array}\right\} \tag{1-99}$$

PROBLEM 1-21

(a) Show that function $\psi = R\Theta\Phi$ corresponding to the lowest value of E_n is

$$\psi = e^{-\alpha r}$$

(b) What are the dimensions of α in this expression? What is its physical meaning? Show that the reciprocal of its value is about 0.5 A.

PROBLEM 1-22 By substituting $R_{3,0} = e^{-\alpha r}L_{3,0}$ into Eq. (1-94), show that we verify Eq. (1-97) for $n = 3$.

Combining the individual solutions of Eqs. (1-95), (1-96) and (1-99), we obtain the expressions for the wave functions as given in Table 1-3.

TABLE 1-3

n	l	m_l	E_n	ψ_{nlm}
1	0	0	$-13.6\ eV$	$e^{-\alpha r}$
2	0	0	-3.4	$(1 - \alpha r)\ e^{-\alpha r}$
2	1	0	-3.4	$re^{-\alpha r} \cos\theta$
2	1	± 1	-3.4	$re^{-\alpha r} \sin\theta\ e^{\pm i\varphi}$
3	0	0	-1.5	$(3 - 6\ \alpha r + 2\alpha^2 r^2)\ e^{-\alpha r}$
3	1	0	-1.5	$(2 - \alpha r)\ re^{-\alpha r} \cos\theta$
3	1	± 1	-1.5	$(2 - \alpha r)\ re^{-\alpha r} \sin\theta\ e^{\pm i\varphi}$
3	2	0	-1.5	$r^2 e^{-\alpha r}\ (3\cos^2\theta - 1)$
3	2	± 1	-1.5	$r^2 e^{-\alpha r} \sin\theta \cos\theta\ e^{\pm i\varphi}$
3	2	± 2	-1.5	$r^2 e^{-\alpha r} \sin^2\theta\ e^{\pm 2i\varphi}$

PROBLEM 1-23 Verify that all possible combinations of quantum numbers for $n \leqq 3$ are shown in Table 1-3.

In Fig. 1-17 we have plotted the probability density, $\psi^*_{nlm}\ \psi_{nlm}$ for a number of the functions listed in the table. The following points should be kept in mind when using this figure:

(1) The probability-densities are all symmetric about the z-axis—that is, they are independent of φ—since $e^{-i\varphi}\ e^{i\varphi} = 1$.

(2) The densities decay as we get farther from the nucleus in accordance with $e^{-\alpha r}$.

(3) For the higher values of n, the functions $R_{n,l}$ show *nodes* (points where the function vanishes), so the densities show corresponding rings.

(4) The individual parts of this figure are not all to the same scale.

PROBLEM 1-24

(a) In solving the Schroedinger equation, we have imposed boundary conditions on the solutions, such as those of Eqs. (1-73) and (1-74). Another condition that is convenient to use is *normalization*, which can be expressed

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \psi^*\psi\ dx\ dy\ dz = 1$$

Recalling that $\psi^*\psi$ is the probability-density, give a physical interpretation of this equation.

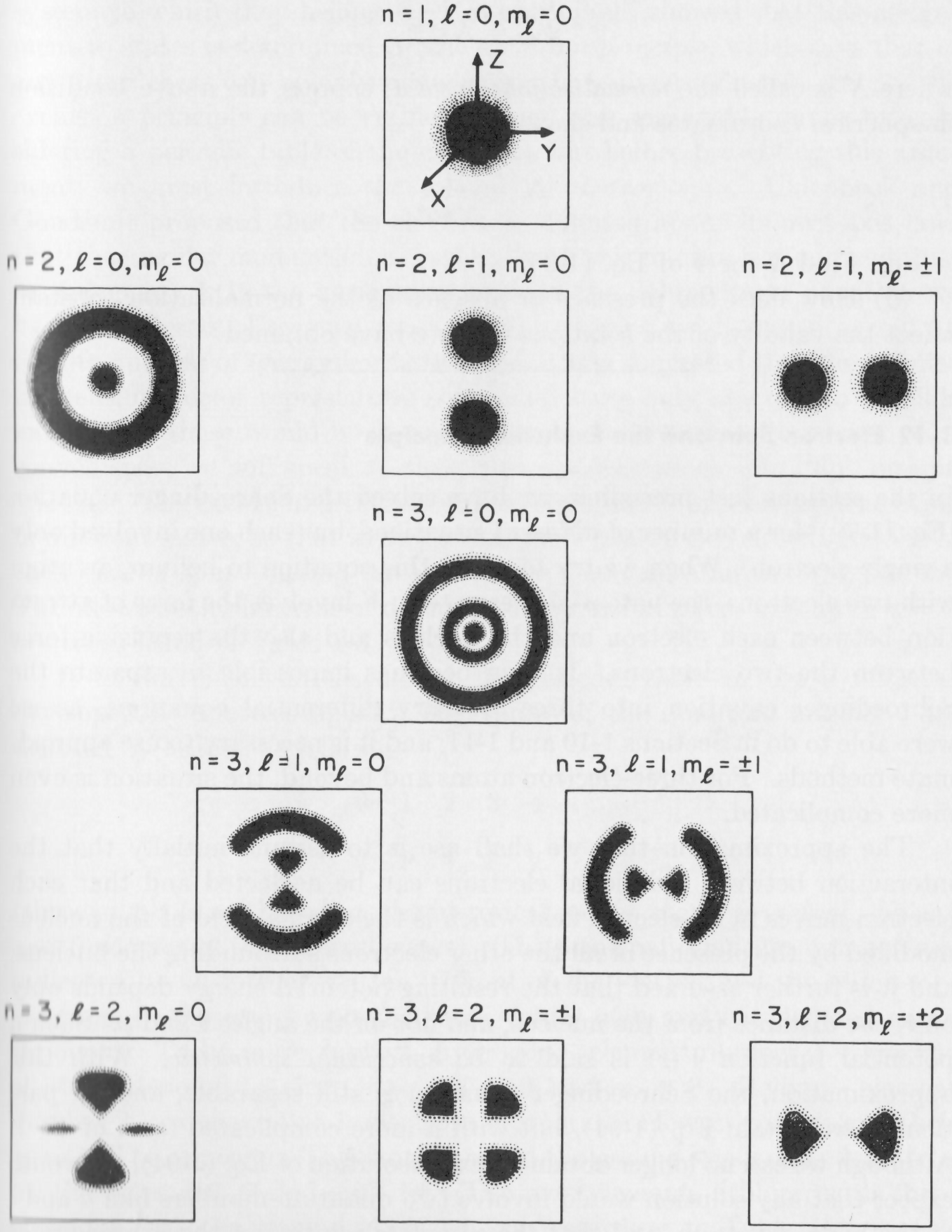


FIG. 1-17. Cross-sections in the XY-plane of the solutions of the Schroedinger equation for the hydrogen atom for n = 1, 2, and 3. The shading represents the exponential decay as a function of r. Note that the scales are not the same for all the functions. (After R. B. Leighton, *Principles of Modern Physics*, McGraw-Hill, New York, 1959, with permission.)

(b) Writing ψ_{100} as

$$\psi_{100} = Ne^{-\alpha r}$$

where N is called the *normalization constant*, express the above condition in spherical coordinates and show that

$$N = \frac{1}{\sqrt{\pi\alpha^3}}$$

(c) Find N for Φ of Eq. (1-95).

(d) How does the presence or absence of the normalization constant affect the validity of the solutions that we have obtained?

1-12 Electron Spin and the Exclusion Principle

In the sections just preceding, we have solved the Schroedinger equation [Eq. (1-79)] for a number of different situations, but each one involved only a single electron. When we try to apply this equation to helium, an atom with two electrons, the potential energy term V involves the force of attraction between each electron and the nucleus and also the repulsive force between the two electrons. It then becomes impossible to separate the Schroedinger equation into three ordinary differential equations, as we were able to do in Sections 1-10 and 1-11, and it is necessary to use approximate methods. For three-electron atoms and beyond, the situation is even more complicated.

The approximation that we shall use is to assume initially that the interaction between individual electrons can be neglected and that each electron moves in an electric field which is the original field of the nucleus modified by the presence of all the other electrons surrounding the nucleus, and it is further assumed that the resulting potential energy depends only on r, the distance from the nucleus, and not on the angles θ and φ. Such a potential function $V(r)$ is said to be *spherically symmetric*. With this approximation, the Schroedinger equation is still separable, and, in particular, we obtain Eq. (1-94), but with a more complicated form of $V(r)$. Although we can no longer obtain an exact solution of Eq. (1-94), we would expect that any solution would involve two quantum numbers like n and l of the hydrogen atom, and that the complete solution of the Schroedinger equation will then involve three quantum numbers. Unlike Eq. (1-94), however, we now expect the total energy to depend on both n and l, instead of n alone.

Calling these quantum numbers n, l, and m_l, as before, we can specify the state of each electron in the atom by giving the associated wave function ψ_{nlm} and energy E_{nl}. It was initially thought that electrons in atoms could occupy states governed only by the amount of energy available in the system to which they belonged. However, Pauli showed that this assignment to states is determined by the *exclusion principle*, which says that in any atom there can be only one electron in a given quantum state. The exclusion principle can be verified beyond any reasonable doubt by considering a periodic table of the elements, but before presenting this argument, we must introduce the concept of *electron spin*. Uhlenbeck and Goudsmit proposed that the electron is spinning about its own axis, and that the angular momentum associated with this motion is quantized, just as the energy is in the various solutions of the Schroedinger equation we have considered. However, most of the examples we discussed permit an infinite number of energy levels, whereas it was suggested that the angular momentum vector representing spin could have only one of two possible values, and these would be of equal magnitude but opposite in sign. For convenience, we will speak of these two possibilities as spin "up" or spin "down." The exclusion principle regards electrons of opposite spin as being in different states, so that in any atom, there can be two electrons – one of each kind of spin – having the same three quantum numbers n, l, and m_l. The allowed combinations of quantum numbers for values of n up to $n = 3$ are those listed in Table 1-3.

It is convenient in discussing the periodic table to use a notation developed by spectroscopists. The values of the quantum number l are denoted by letters, as follows:

$$\begin{array}{cccccc} 0 & 1 & 2 & 3 & 4 & \ldots \\ s & p & d & f & g & \ldots \end{array}$$

(these symbols are based on the historical names *sharp*, *principal*, *diffuse*, and *fundamental* for spectral series). The principal quantum number n is indicated by an integer written in front of the l-letter, and the number of electrons in the atom with the same values of n and l is denoted as an exponent. To be more explicit, hydrogen – element number 1 – has one electron, denoted $1s$, since $n = 1$, $l = 0$, and $m_l = 0$. Helium – element number 2 – can have two $1s$ electrons in its state of lowest energy, and their spins will be opposite. This state is called the *ground state* and is denoted – by the rules just stated – as $1s^2$. The next element, lithium, with three electrons, has the ground state $1s^2\ 2s$. According to Table 1-3, two $2s$ electrons and six $2p$ electrons are allowed, so that element number 10, neon, has a total of eight electrons with quantum number $n = 2$ and its ground state is denoted $1s^2\ 2s^2\ 2p^6$.

The electrons belonging to a given value of n are said to form *shells* and when an atom has the maximum number of electrons for each n value, the shell is *closed*. Table 1-5 lists the elements with atomic numbers 1 through

54 (this group is of primary interest in semiconductor physics) and shows their electronic configuration as determined by spectroscopic methods. It should be observed that the shells fill up in approximately the order 1*s*, 2*s*, 2*p*, 3*s*, 3*p*, 4*s*, 3*d*, 4*p*, 5*s*, 4*d*, and 5*p*. This table enables us to locate atoms whose outermost shells are similar. For example, the inert gases shown are helium, with a filled shell corresponding to $n = 1$; neon, with a filled $n = 2$ shell; argon, with a filled $n = 3$ shell; krypton, with a filled 4*p* *subshell;* and xenon, with a filled 5*p* subshell. These four elements illustrate how the exclusion principle can account for the repetition of chemical properties in the periodic table. In a later chapter we shall consider in detail the periodic series shown in Table 1-4. Carbon (as diamond), silicon, germanium, and tin (in the gray modification) all have the same crystal structure; Table 1-4 shows that they have a similar electronic configuration. That is, they all have four electrons in their last, unfilled shell. These electrons are called the *valence electrons* and are responsible for the binding of the atoms in the crystal.

TABLE 1-4

Element	*Atomic number*	*Ground state*
Carbon	6	$1s^2\,2s^2\,2p^2$
Silicon	14	$1s^2\,2s^2\,2p^6\,3s^2\,3p^2$
Germanium	32	$1s^2\,2s^2\,2p^6\,3s^2\,3p^6\,3d^{10}\,4s^2\,4p^2$
Tin	50	$1s^2\,2s^2\,2p^6\,3s^2\,3p^6\,3d^{10}\,4s^2\,4p^6\,4d^{10}\,5s^2\,5p^2$

1-13 Electronic Energy Bands

The final application of the Schroedinger equation we shall consider in this chapter is to an electron moving in an idealized solid body. A hydrogen atom has the potential energy given by Eq. (1-89) and plotted in Fig. 1-18.

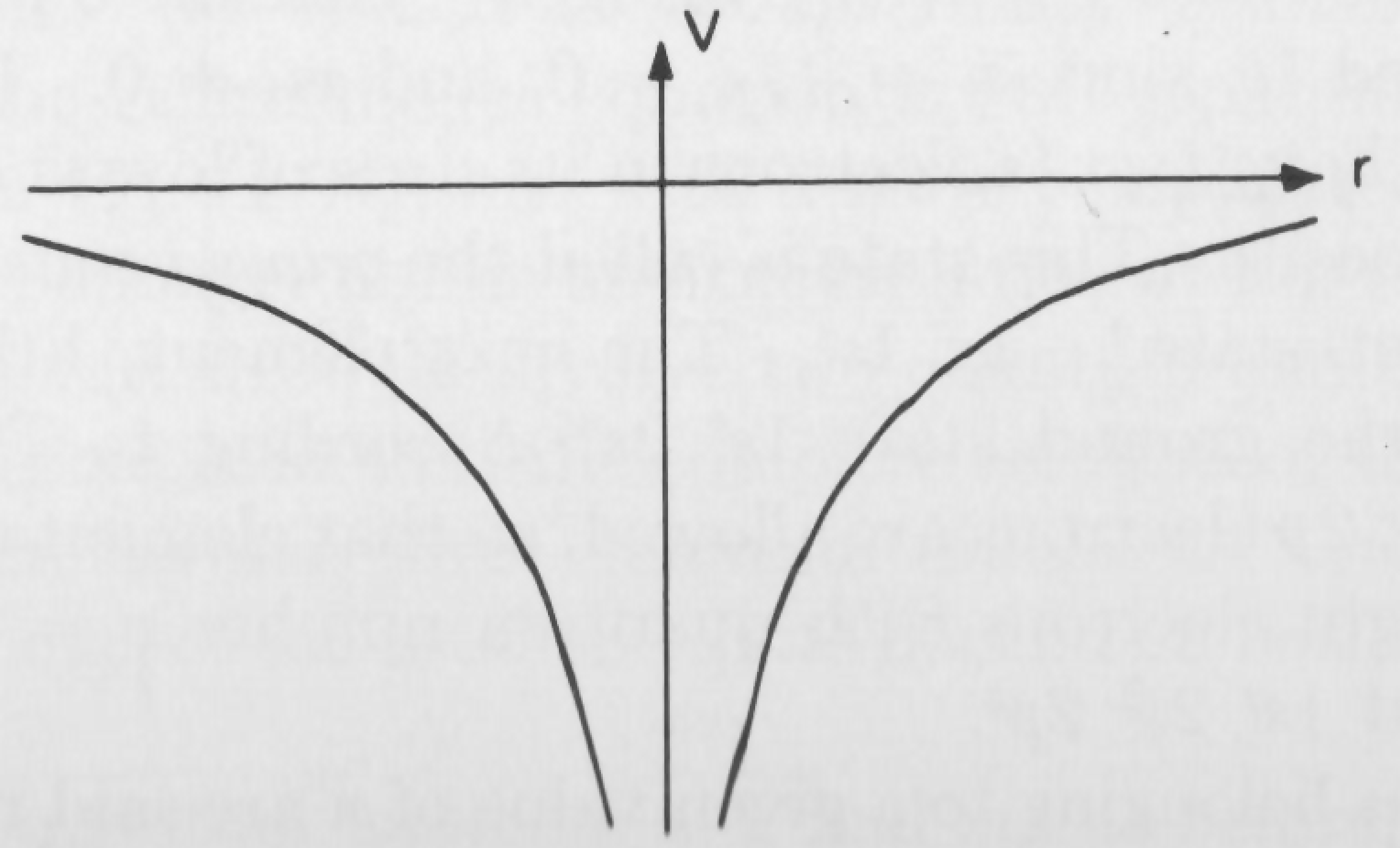


FIG. 1-18. Potential energy *V* as a function of distance *r* for the hydrogen atom. (From Sproull, *op. cit.*)

TABLE 1-5

Atomic number	*Element*	*Electron configuration*										
		1*s*	2*s*	2*p*	3*s*	3*p*	3*d*	4*s*	4*p*	4*d*	5*s*	5*p*
1	H	1										
2	He	2										
3	Li	2	1									
4	Be	2	2									
5	B	2	2	1								
6	C	2	2	2								
7	N	2	2	3								
8	O	2	2	4								
9	F	2	2	5								
10	Ne	2	2	6								
11	Na	2	2	6	1							
12	Mg	2	2	6	2							
13	Al	2	2	6	2	1						
14	Si	2	2	6	2	2						
15	P	2	2	6	2	3						
16	S	2	2	6	2	4						
17	Cl	2	2	6	2	5						
18	Ar	2	2	6	2	6						
19	K	2	2	6	2	6		1				
20	Ca	2	2	6	2	6		2				
21	Sc	2	2	6	2	6	1	2				
22	Ti	2	2	6	2	6	2	2				
23	V	2	2	6	2	6	3	2				
24	Cr	2	2	6	2	6	5	1				
25	Mn	2	2	6	2	6	5	2				
26	Fe	2	2	6	2	6	6	2				
27	Co	2	2	6	2	6	7	2				
28	Ni	2	2	6	2	6	8	2				
29	Cu	2	2	6	2	6	10	1				
30	Zn	2	2	6	2	6	10	2				
31	Ga	2	2	6	2	6	10	2	1			
32	Ge	2	2	6	2	6	10	2	2			
33	As	2	2	6	2	6	10	2	3			
34	Se	2	2	6	2	6	10	2	4			
35	Br	2	2	6	2	6	10	2	5			
36	Kr	2	2	6	2	6	10	2	6			
37	Rb	2	2	6	2	6	10	2	6		1	
38	Sr	2	2	6	2	6	10	2	6		2	
39	Y	2	2	6	2	6	10	2	6	1	2	
40	Zr	2	2	6	2	6	10	2	6	2	2	
41	Nb	2	2	6	2	6	10	2	6	4	1	
42	Mo	2	2	6	2	6	10	2	6	5	1	
43	Tc	2	2	6	2	6	10	2	6	6	1	
44	Ru	2	2	6	2	6	10	2	6	7	1	
45	Rh	2	2	6	2	6	10	2	6	8	1	
46	Pd	2	2	6	2	6	10	2	6	10		
47	Ag	2	2	6	2	6	10	2	6	10	1	
48	Cd	2	2	6	2	6	10	2	6	10	2	
49	In	2	2	6	2	6	10	2	6	10	2	1
50	Sn	2	2	6	2	6	10	2	6	10	2	2
51	Sb	2	2	6	2	6	10	2	6	10	2	3
52	Te	2	2	6	2	6	10	2	6	10	2	4
53	I	2	2	6	2	6	10	2	6	10	2	5
54	Xe	2	2	6	2	6	10	2	6	10	2	6

A *crystal* consists of a collection of atoms arranged in a regular array. If it were possible to make a one-dimensional hydrogen crystal, the potential energy as a function of distance would be expected to appear as shown in Fig. 1-19. The separation L between atoms is called the *lattice constant.*

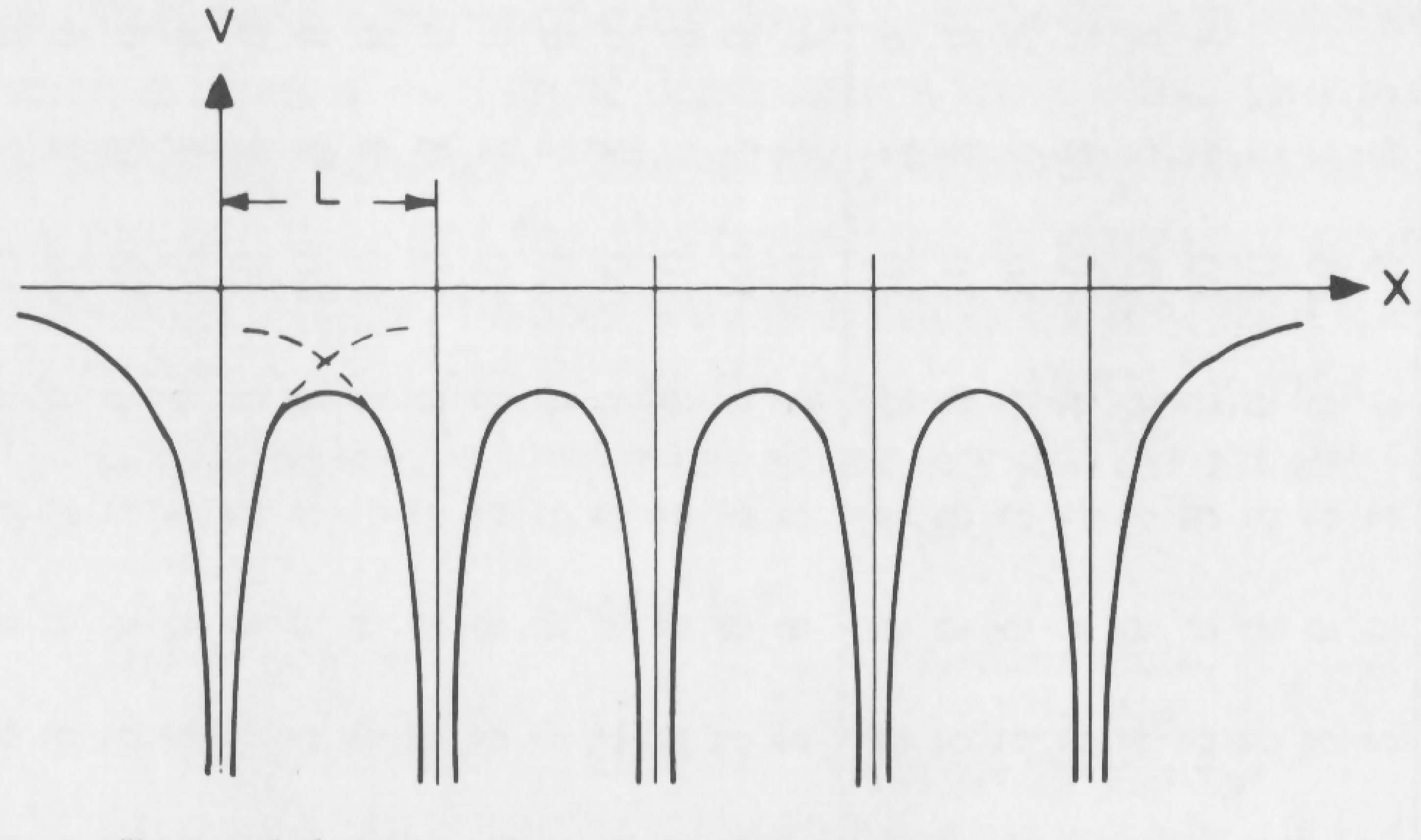


FIG. 1-19. Potential energy for a hypothetical one-dimensional hydrogen crystal. (From Sproull, *op. cit.*)

For mathematical convenience, we first replace the potential energy curve of Fig. 1-19 with an idealized curve of rectangular shape and then shift its position with respect to both the x-axis and the V-axis, as shown in Fig. 1-20. We now have two Schroedinger equations like Eqs. (1-67) and (1-66) to solve

$$\frac{d^2\psi}{dx^2} + \alpha^2\psi = 0 \quad \text{for } 0 < x < a \tag{1-100}$$

$$\frac{d^2\psi}{dx^2} - \beta^2\psi = 0 \quad \text{for } -b < x < 0 \tag{1-101}$$

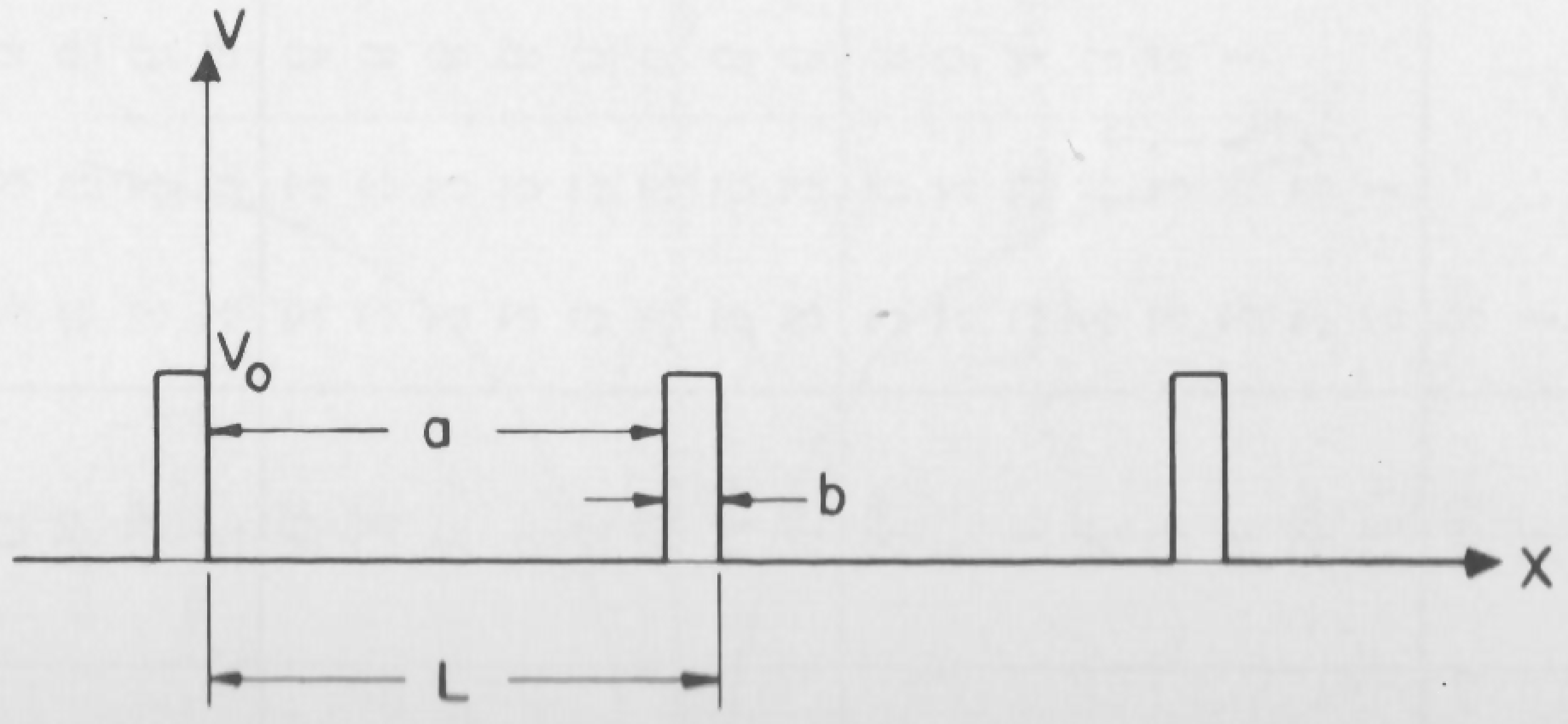


FIG. 1-20. Simplified potential energy for a one-dimensional crystal. (From Sproull, *op. cit.*)

Since V is *periodic*, repeating itself as x goes to $x + L$, $x + 2L$, etc., we will try a solution

$$\psi(x) = u(x)\, e^{ikx} \tag{1-102}$$

where $u(x)$ will be required to be periodic. If we compare this with Eq. (1-50), with the time chosen as $t = 0$ for convenience,

$$y = ae^{ikx} \tag{1-103}$$

we see that in using Eq. (1-102) we are assuming that the electron is a wave with wave length $\lambda = 2\pi/k$ and amplitude $u(x)$, which varies periodically as we go from one atom to the next. This means we are saying that the electrons in this crystal do not belong to individual atoms; rather, they are equivalent to waves spread through the whole length of the lattice. These waves are called *Bloch waves.*

Substituting Eq. (1-102) into Eqs. (1-100) and (1-101), we obtain

$$d^2u/dx^2 + 2ik\, du/dx + (\alpha^2 - k^2)u = 0 \tag{1-104}$$

$$d^2u/dx^2 + 2ik\, du/dx - (\beta^2 + k^2)u = 0 \tag{1-105}$$

The solutions to these equations can be verified to be

$$u_1 = Ae^{i(\alpha-k)x} + Be^{-i(\alpha+k)x} \tag{1-106}$$

$$u_2 = Ce^{(\beta-ik)x} + De^{-(\beta+ik)x} \tag{1-107}$$

Using the same arguments which led to Eqs. (1-73) and (1-74), at $x = 0$ we impose the boundary conditions:

$$u_1(0) = u_2(0),\ du_1(0)/dx = du_2(0)/dx \tag{1-108}$$

Since $u(x)$ is required to be periodic, we have, in addition,

$$u_1(a) = u_2(-b),\ du_1(a)/dx = du_2(-b)/dx \tag{1-109}$$

Inserting Eqs. (1-106) and (1-107) into Eqs. (1-108) and (1-109) gives the following four equations for A, B, C, and D:

$$\left.\begin{aligned} A + B &= C + D \\ i(\alpha - k)A - i(\alpha + k)B &= (\beta - ik)C - (\beta + ik)D \\ e^{i(\alpha-k)a}A + e^{-i(\alpha+k)a}B &= e^{-(\beta-ik)b}C + e^{(\beta+ik)b}D \\ i(\alpha - k)e^{i(\alpha-k)a}A - i(\alpha - k)e^{-i(\alpha+k)a}B &= (\beta - ik)e^{-(\beta-ik)b}C \\ &\quad -(\beta + ik)e^{(\beta+ik)b}D \end{aligned}\right\} \tag{1-110}$$

These equations are *homogeneous* (they contain no terms not involving A, B, C, or D), so they will have a solution, provided the determinant of the coefficients of A, B, C, and D vanishes. Expanding this determinant gives

$$[(\beta^2 - \alpha^2)/2\alpha\beta] \sinh \beta b \sin \alpha a + \cosh \beta b \cos \alpha a = \cos k(a + b) \tag{1-111}$$

where the hyperbolic functions are defined by

$$\left.\begin{aligned}\sinh \beta b &= (e^{\beta b} - e^{-\beta b})/2\\ \cosh \beta b &= (e^{\beta b} + e^{-\beta b})/2\end{aligned}\right\} \tag{1-112}$$

All quantities involved in Eq. (1-111) except E (through α and β) and k are already determined. Since $\cos kL$ on the right-hand side can only assume values from -1 to 1, it follows that a value for E exists only when the left-hand side, which we denote by $L(E)$, lies in the same range. Figure 1-21a is a plot of the left-hand side of Eq. (1-111) as a function of E with V_0 chosen to have the same numerical value as in Section 1-14

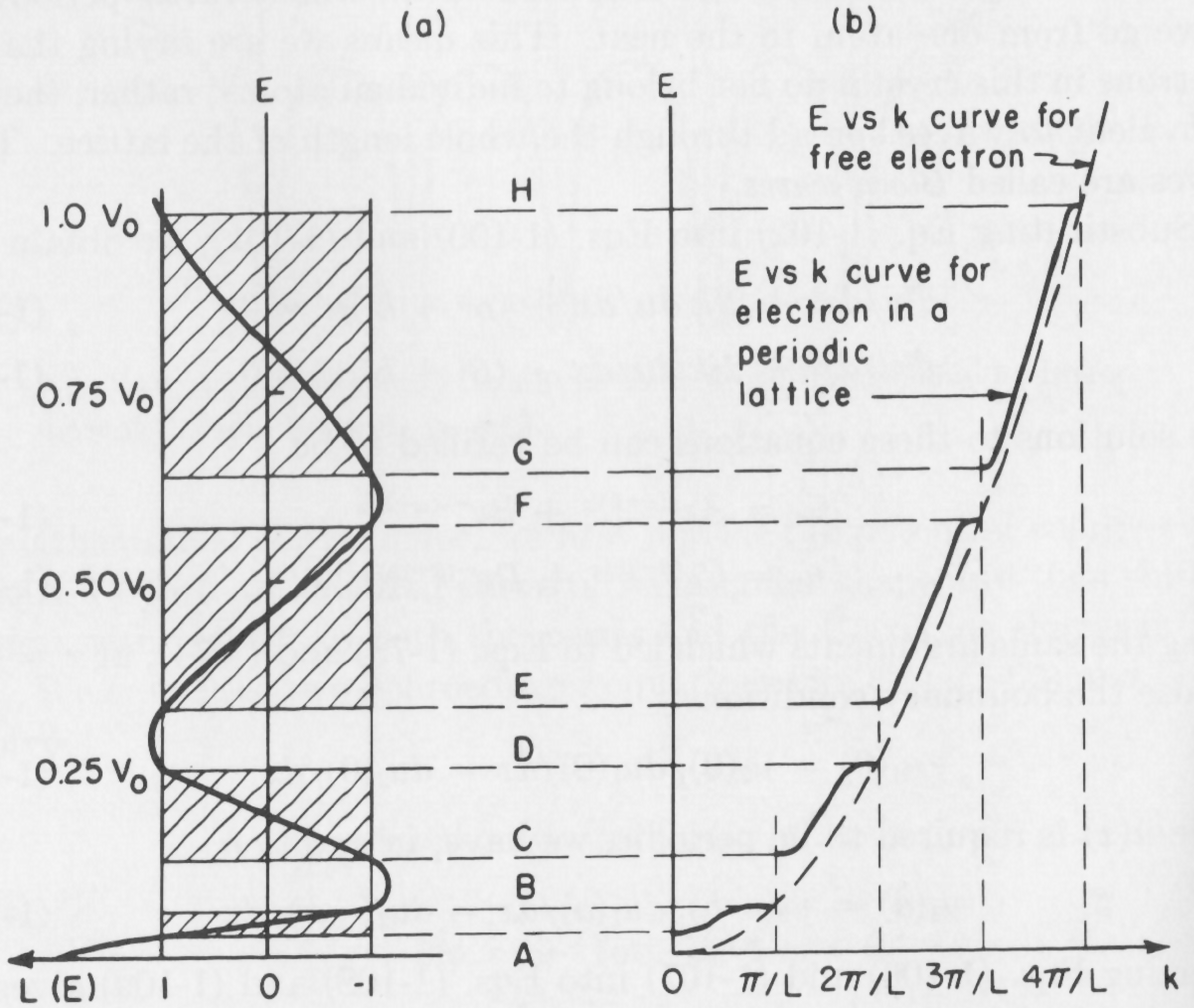


FIG. 1-21. (a) Allowed and forbidden energy bands for a one-dimensional crystal with $b/a = 1/24$. (After Rojansky, *op. cit.*) (b) Variation of energy E with propagation constant k. (From Sproull, *op. cit.*)

$$2m\,V_0 w^2/\hbar^2 = 2mV_0 a^2/4\hbar^2 = 36$$

and with $b/a = 1/24$. The shaded areas of the graph show the ranges of E for which values of E exist. These ranges are called *allowed energy bands*. The unshaded regions are called *forbidden energy bands*. According to the figure, the allowed bands lie in the approximate ranges shown in Table 1-6, as given by Rojansky (1938).[4] We have also put down for comparison the four energy levels of Section 1-9, and we see that the bands represent a spreading of the levels as we go from a single well to a periodic array.

TABLE 1-6

Periodic array of wells	*Single well*
$0.03\ V_0 \leq E \leq 0.06\ V_0$	$0.05\ V_0$
$0.13\ V_0 \leq E \leq 0.25\ V_0$	$0.20\ V_0$
$0.33\ V_0 \leq E \leq 0.57\ V_0$	$0.45\ V_0$
$0.64\ V_0 \leq E \leq 0.99\ V_0$	$0.75\ V_0$

We can get some further information of importance from Eq. (1-111) by plotting E vs. k. Since $\cos kL = 1$ when $k = 0$, and $\cos kL = -1$ when $k = \pi/L$, then as k takes on values from 0 to π/L, we can solve Eq. (1-111) for the corresponding value of E, obtaining the graph of Fig. 1-21b. Note that as k passes through π/L there is a discontinuity in E which corresponds to a forbidden band of Fig. 1-21a, and that as k increases, we generate the entire band structure. The energy of an electron moving freely in a vacuum, which we refer to as a *free electron*, is, by Eq. (1-48),

$$E = p^2/2m = \hbar^2 k^2/2m$$

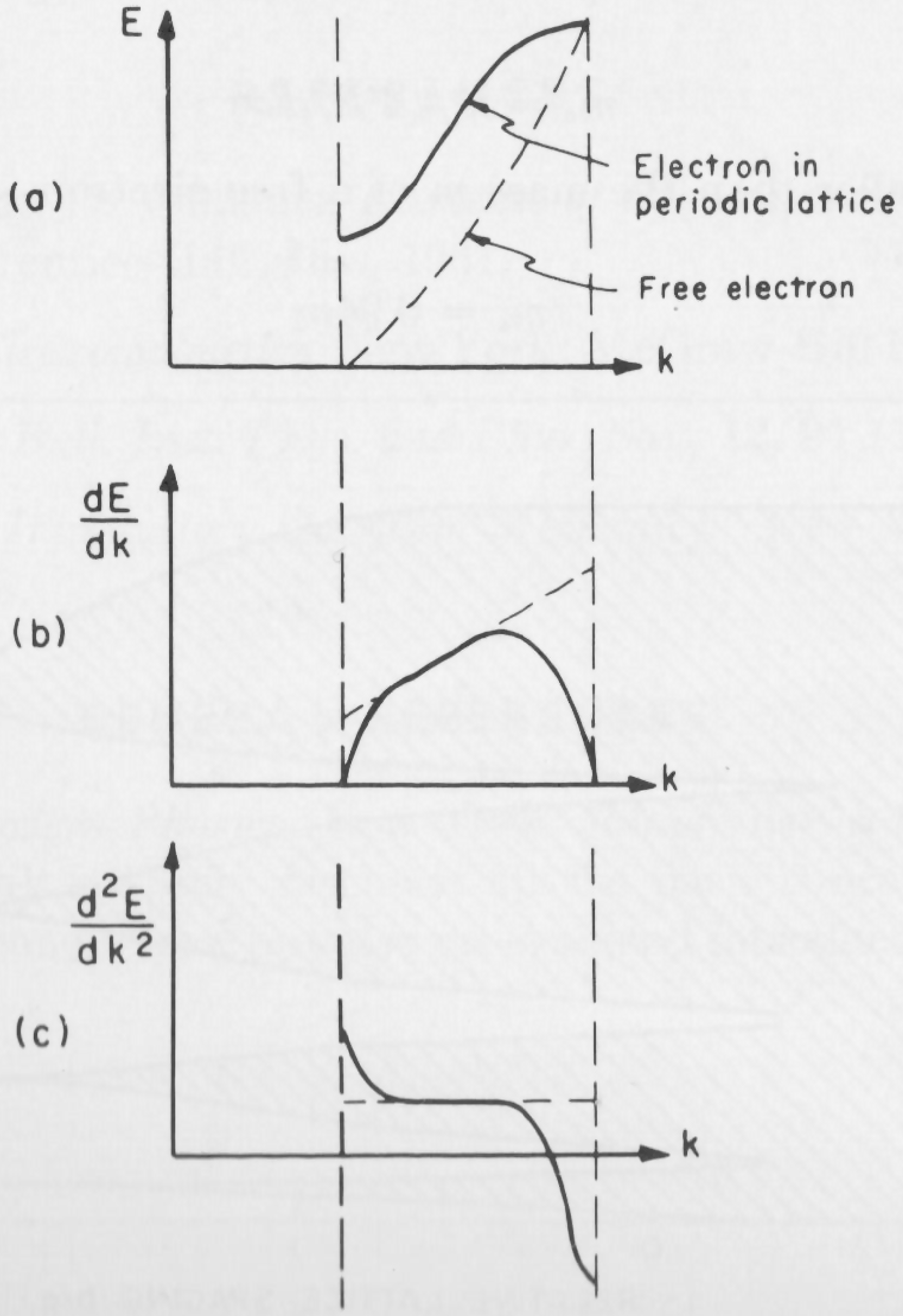


FIG. 1-22. (a) The section of Fig. 1-21(b) lying between levels C and D. (b) The derivative of this curve. (c) The second derivative.

The curve of E vs. k for the free electron is, therefore, a parabola and is shown dotted in Fig. 1-21b. Differentiating this, we obtain

$$dE/dk = \hbar^2 k/m \tag{1-113}$$

and, again,

$$d^2E/dk^2 = \hbar^2/m \tag{1-114}$$

or

$$m = \hbar^2/(d^2E/dk^2) \tag{1-115}$$

Equation (1-115) tells us that the mass of a free electron is inversely proportional to d^2E/dk^2. We note that the solid curves of Fig. 1-21b are roughly parallel to the dotted parabola in the regions which are far from the energy discontinuities. In Fig. 1-22, the graph (a) shows the section of Fig. 1-21b lying between energy levels C and D, and the graphs (b) and (c) directly beneath represent sketches of the approximate behavior of the corresponding first and second derivatives. We note, for example, that, by Eq. (1-113), dE/dk should be a straight line of constant slope for a free electron and d^2E/dk^2 is a constant. The corresponding curve for the electron in the periodic lattice shows that d^2E/dk^2 is larger at the bottom of this band than it is for a free electron. Hence, we define the *effective mass* m_e of an electron in a periodic lattice, in analogy with Eq. (1-115), as

$$m_e = \hbar^2/(d^2E/dk^2) \tag{1-116}$$

Then m_e is smaller than the mass m of a free electron. For potassium, a typical value is

$$m_e = 0.94m$$

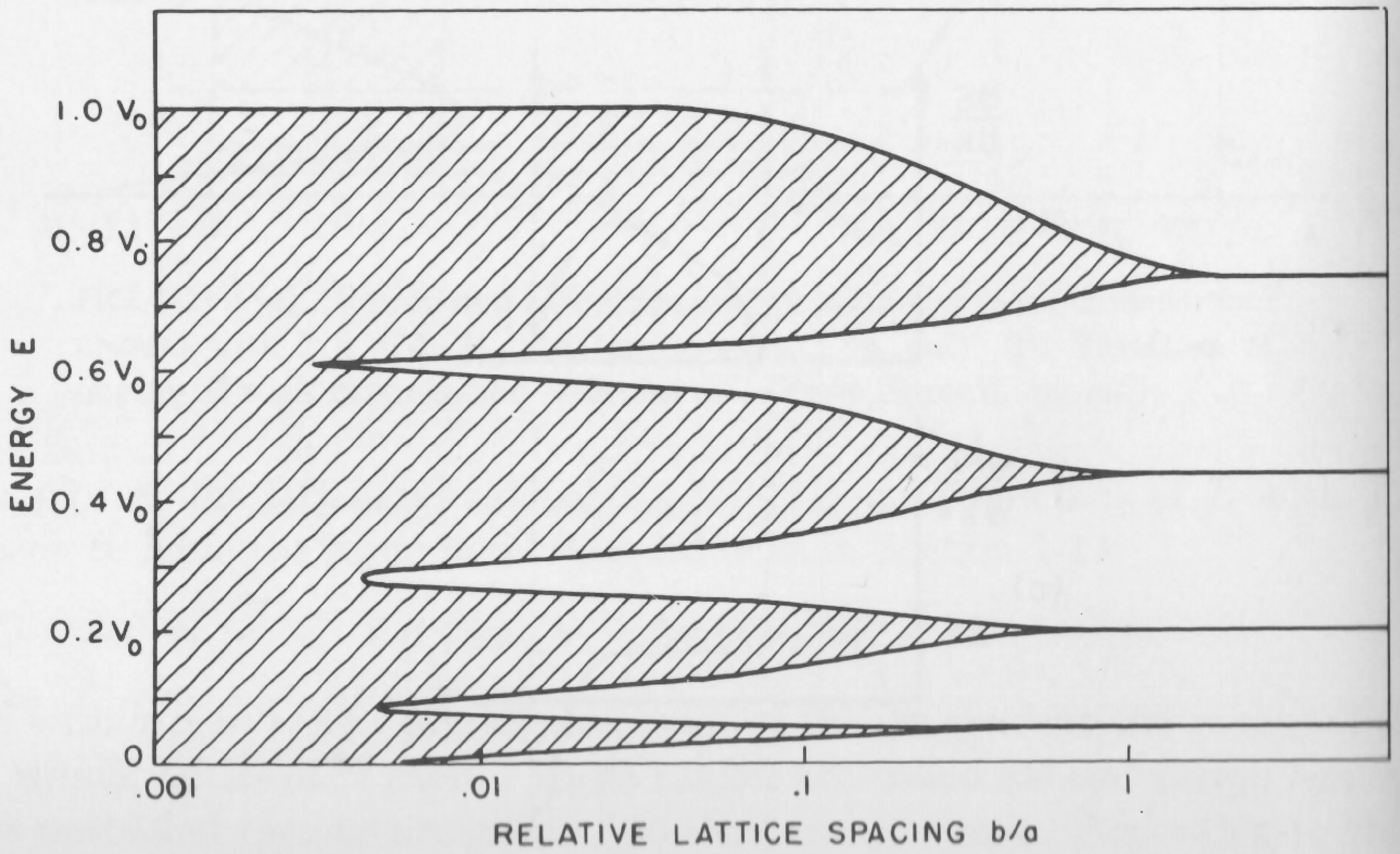


FIG. 1-23. Energy band width as a function of lattice spacing for a one-dimensional crystal.

PROBLEM 1-25

(a) Show by the use of Eq. (1-111) that when $b/a = 10$, the energy bands of Fig. 1-21 reduce to the four discrete levels of Section 1-9.

(b) Show that when $b/a = 1/240$, the bands spread out and completely overlap.

The results of this problem, as well as the ranges of the bands for intermediate values of b/a, are shown in Fig. 1-23. This figure is essentially a diagram showing the width of the allowed and forbidden bands for a one-dimensional crystal as a function of lattice spacing. Note that b/a is plotted logarithmically, owing to the wide range of values we are covering. This figure illustrates the remark made above concerning the broadening of the energy levels as we make the transition from the single well to the periodic array of wells. Figures 1-21a, 1-21b, and 1-23 are three different types of energy band diagrams, all of which are useful in semiconductor research and applications. Later in this book, we shall consider such diagrams for real crystals, and show how they are obtained and used.

REFERENCES

1. Shortley, G. and D. Williams, *Elements of Physics*, 3rd ed. Englewood Cliffs, N. J.: Prentice-Hall, Inc., 1961.
2. Kraus, J. D., *Electromagnetics*. New York: McGraw-Hill Book Co., 1953.
3. Pearson, F. J., *Bull. Inst. Phys. and Phys. Soc.*, **12,** 91 (1961).
4. Rojansky, V., *Introductory Quantum Mechanics*. New York: Prentice-Hall, Inc., 1938.

GENERAL REFERENCES

Sproull, R. L., *Modern Physics*. New York: John Wiley & Sons, 1956.
This excellent book not only discusses all the basic concepts needed for the study of semiconductors, but also gives a good introduction to the idea of energy bands.

2

THERMOSTATICS AND STATISTICAL PHYSICS

2-1 Temperature, Heat, and Internal Energy

The study of problems involving heat has traditionally been called *thermodynamics*. Among the topics covered by thermodynamics is the behavior of chemical systems. If we are dealing with systems whose chemical composition is uniform and whose temperature is everywhere the same, it is more correct to say we are concerned with *thermostatics*.

It was mentioned in Section 1-2 that the fundamental quantity which must be introduced into physics in order to study the flow of heat and its effects is temperature. Two temperature scales are in use in semiconductor research, the *centigrade* (C) and the *Kelvin* (K), defined as follows:

melting point of ice at 1 atm pressure = 0°C = 273.2°K

boiling point of water at 1 atm pressure = 100°C = 373.2°K.

We know that *heat* Q is a form of energy. To be precise, heat is energy which is transferred by virtue of a temperature difference. However, once the transfer has been completed, the energy can no longer be classified as heat, and it is incorrect to talk about the *heat content* of a body. To use an excellent example of Zemansky (1937),[1] during a storm, rain will fall into a lake. After the rain stops, the water which entered the lake is no longer rain — it is water like that originally there. Water is rain only during its passage from sky to lake. It is proper, however, to talk about the gain or loss of energy in a body. Let us take some particular condition of the body as the reference state. Then the energy which it has in any other state with respect to this reference state is called the *internal energy* U of the body. The internal energy can be changed by the flow of heat or by doing work. Let:

$+dQ$ = amount of heat which *enters* a body

$+dW$ = amount of work done *by* a body

with $-dQ$ and $-dW$ representing, repectively, heat which leaves the body and work done on it. Then the principle of conservation of energy requires that:

$$dU = dQ - dW \tag{2-1}$$

Equation (2-1) is known as the *first law of thermodynamics*.

2-2 Entropy

The variables of thermostatics can be divided into two groups. One is the group of *extensive* quantities and these depend upon the mass of material present. Examples are volume V, internal energy U, and number of molecules N. The other group, the *intensive* variables, do not depend on the mass present. Examples are temperature T and pressure P.

We now postulate that there exist states of systems which can be completely characterized by making laboratory measurements of a small number of extensive variables. Such states are called *equilibrium states*.

In Eq. (2-1) we have expressed mathematically the concept of energy conservation as it applies in thermostatics. Although this principle is fundamental, it is not sufficient to describe the behavior of thermal systems. To give two examples of Sears (1950),[2] we never find in nature that a body at a uniform temperature will spontaneously get hotter or colder in one section, or that a chemical solution will separate into regions of different concentration. Since energy would be conserved in these examples, some other variable must be responsible for what we consider normal behavior. This variable is called *entropy* S. We will define entropy by imposing the following conditions on it, and later will look into its physical meaning:

(1) Entropy is an extensive variable depending upon U, V, and N, and the total entropy of a group of systems is the sum of the individual entropies.

(2) Entropy is defined only for equilibrium states.

(3) The entropy of an unconstrained system tends to go to a maximum, where *unconstrained* means that all differences in temperature, concentration, etc. are allowed to equalize themselves.

Mathematically, condition (1) for a system with two chemical constituents is

$$S = S(U, V, N_1, N_2) \tag{2-2}$$

Equation (2-2) can be solved for U to give

$$U = U(S, V, N_1, N_2) \tag{2-3}$$

2-3 Temperature, Pressure, and Chemical Potential

Taking the differential of U in Eq. (2-2) gives

$$dU = \left.\frac{\partial U}{\partial S}\right)_{V,N_1,N_2} dS + \left.\frac{\partial U}{\partial V}\right)_{S,N_1,N_2} dV + \left.\frac{\partial U}{\partial N_1}\right)_{S,V,N_2} dN_1 + \left.\frac{\partial U}{\partial N_2}\right)_{S,V,N_1} dN_2 \tag{2-4}$$

The partial derivatives in this equation have names, as follows:

$$\left.\frac{\partial U}{\partial S}\right)_{V,N_1,N_2} \equiv T \equiv \textit{absolute temperature} \tag{2-5}$$

$$-\left.\frac{\partial U}{\partial V}\right)_{S,N_1,N_2} \equiv P \equiv \textit{pressure} \tag{2-6}$$

$$\left.\frac{\partial U}{\partial N_1}\right)_{S,V,N_2} \equiv K_1 \equiv \textit{chemical potential of constituent number 1} \tag{2-7}$$

Then Eq. (2-4) becomes

$$dU = T\,dS - P\,dV + K_1\,dN_1 + K_2\,dN_2 \tag{2-8}$$

Equation (2-5) is the mathematical definition of entropy in terms of the fundamental quantity T and the derived quantity U, and it shows us that the units of entropy are joules/degree. We note that, although temperature is defined in terms of entropy by Eq. (2-5), we choose the absolute (or Kelvin) degree as the fundamental unit as a matter of convenience.

If we keep the amount of each of the two constituents constant, so that dN_1 and dN_2 are zero, then Eq. (2-8) becomes

$$dU = T\,dS - P\,dV \tag{2-9}$$

This will agree with Eq. (2-1) if

$$dW = P\,dV \tag{2-10}$$

and

$$dQ = T\,dS \tag{2-11}$$

Equation (2-10) is the same as Eq. (1-11), showing why we identify $\partial U/\partial V)_{S,N}$ as the negative of the pressure.

Now consider two simple systems connected by a solid, immovable wall which conducts heat. If U_1 is the internal energy of system 1 and U_2 is the internal energy of system 2, then

$$U_1 + U_2 = \text{constant}$$

if the energy is conserved, and

$$dU_1 + dU_2 = 0 \tag{2-12}$$

If we use the maximum entropy principle discussed in the previous section,

$$dS = 0$$

By our definition of S

$$dS = dS_1 + dS_2 = 0 \tag{2-13}$$

Since the wall does not permit changes in N or V, by Eqs. (2-2) and (2-5),

$$dS_1 = \frac{\partial S_1}{\partial U_1}\,dU_1 = \frac{1}{T_1}\,dU_1 \tag{2-14}$$

and

$$dS_2 = \frac{1}{T_2}\,dU_2 \tag{2-15}$$

Combining these equations yields

$$T_1 = T_2 \tag{2-16}$$

This is the condition for *thermal equilibrium*, and is seen to be a consequence of the properties of entropy.

PROBLEM 2-1 Let the wall now be movable, so that the volume of each system can vary, although the total volume is fixed. Show that the conditions for *mechanical equilibrium* are

$$T_1 = T_2 \tag{2-16}$$

$$P_1/T_1 = P_2/T_2 \tag{2-17}$$

Now make the wall immovable again, but allow it to be permeable to one of the constituents of the two systems. Then, as before:

$$dU_1 = -dU_2$$

and, similarly, an increase in amount of material in one system is accompanied by an equal decrease in the other, or

$$dN_1 = -dN_2$$

Using the same reasoning as before, we obtain the conditions for *matter-flow equilibrium*,

$$T_1 = T_2 \tag{2-16}$$

and

$$K_1 = K_2 \tag{2-18}$$

Now Eq. (2-8) shows that terms of the form $K\,dN$ can be thought of as "chemical work," and, by analogy with $P\,dV$, we can think of the chemical potential as the driving "pressure" producing a flow of matter. This flow ceases when the chemical potentials in the two systems are equalized, as indicated by the condition $K_1 = K_2$.

2-4 Reversible and Irreversible Processes; the Meaning of Entropy

All real thermal processes are irreversible. We define an irreversible process as one which is initiated by the removal of a constraint, but which cannot be reversed by the removal of a second constraint. To make this idea clearer, consider a well-insulated container divided into two parts by a removable, insulated partition (Fig. 2-1). Into one compartment put 1 kg of water at 0°C; into the other put 1 kg of water at 100°C. Now suppose we remove the partition; this is what is meant by removing a constraint. After a certain length of time we will have 2 kg of water at 50°C. There is no way in which we can remove a second constraint and return the compartments to their original temperatures. Further, since we stipulated in Section 2-2 that the entropy of an unconstrained system goes to a maximum, we know that the entropy of the mixture is higher than that of the initial system. The necessity for the entropy to increase during a real process may be thought of in an intuitive way as the driving "force" which causes the two 1-kg masses of water to equalize their temperature and prevents the reverse process from occurring by the simple removal of another constraint. Now we can imagine an ideal process in which the entropy remains constant and define this as a *reversible* process. To clarify the difference between reversible and irreversible processes, let us compute the entropy changes associated with this example.

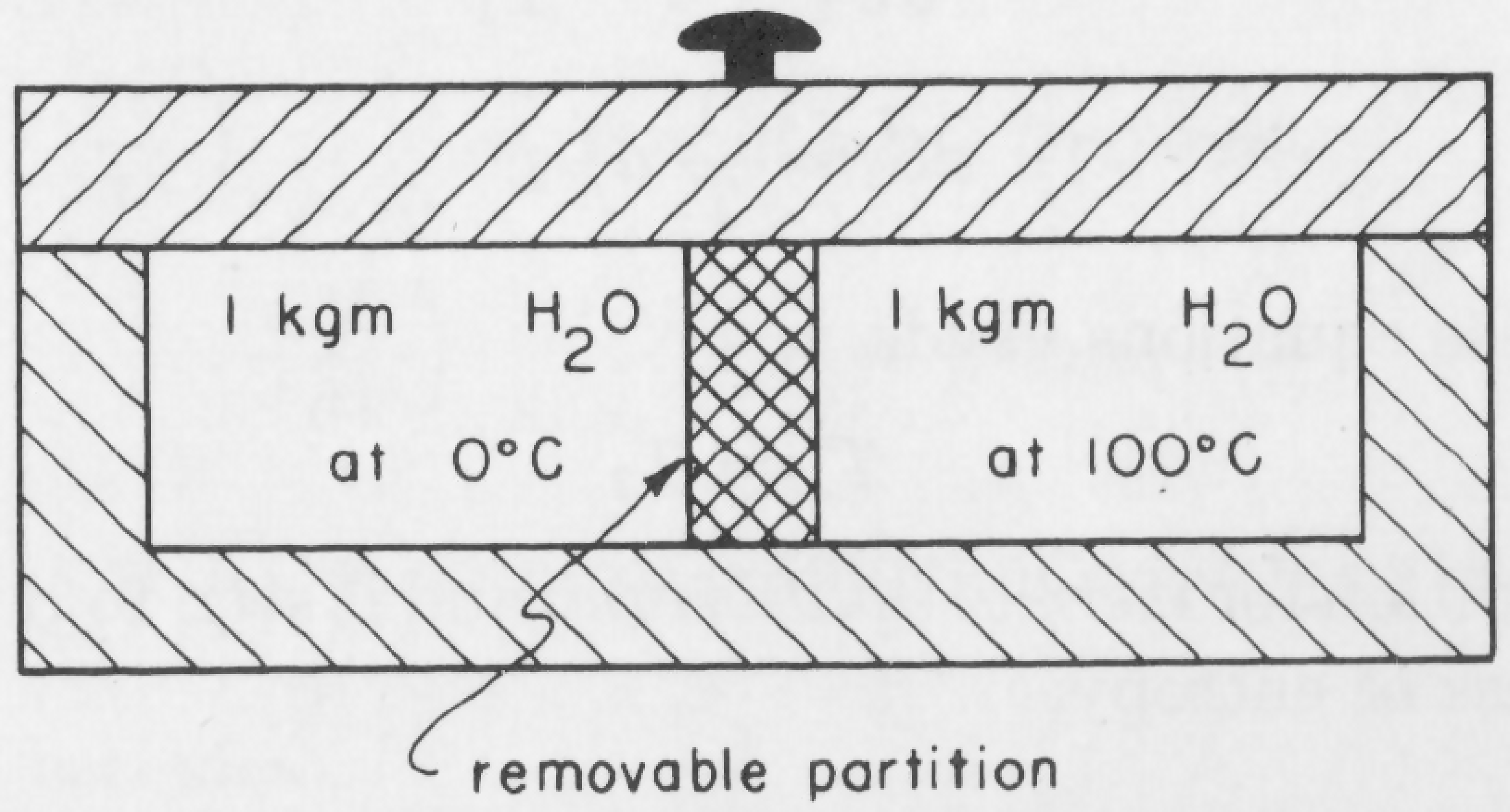


FIG. 2-1. A system used to illustrate the change in entropy associated with mixing.

When heat flows into a body, the temperature generally rises. Define the *specific heat capacity at constant pressure* c_p of a substance as the ratio of heat dQ per unit mass m flowing into it to the corresponding rise dT in temperature. Then

$$dQ = mc_p\,dT$$

and, by Eq. (2-11),

$$dS = mc_p\,dT/T$$

For water

$$c_p = 4.18 \times 10^3 \text{ joules/kg-deg}$$

If the temperature of a substance goes from T_1 to T_2, the entropy change is then

$$\int_{S_1}^{S_2} dS = \int_{T_1}^{T_2} mc_p\,dT/T$$

or

$$S_2 - S_1 = mc_p \ln (T_2/T_1) \tag{2-19}$$

if we use the reasonable approximation that c_p is independent of temperature. For the water originally at 0°C,

$$S_2 - S_1 = 1 \times 4.18 \times 10^3 \times \ln (323/273)$$
$$= 704 \text{ joules/deg}$$

[Note that we use the absolute scale which the definition [Eq. (2-5)] requires.] For the other compartment,

$$S_2 - S_1 = 1 \times 4.18 \times 10^3 \times \ln (323/373)$$
$$= -603 \text{ joules/deg}$$

Thus, the entropy in the compartment which started at 0°C increased by 704 joules/deg, the other compartment had a decrease of 603 joules/deg, and the entropy of the complete system increased, as we would expect for an irreversible process, by 101 joules/deg. Next, let us go through the same process in a reversible fashion. In order to do this, we must imagine we have available an infinite number of *heat reservoirs*, ranging in temperature from 273°K to 373°K in infinitesimal steps. A heat reservoir is a body whose heat capacity is so large that it can accept or give up heat without changing its temperature. If the water at 273°K is placed in contact with the heat reservoir at $(273 + dT)°$, heat will flow from the reservoir to the water so slowly that at any given time we can specify definite values for the extensive variables, thus satisfying the definition of equilibrium. By moving the water from one reservoir to another, we can bring it up to 323°K. The change in entropy of the water, by Eq. (2-19), depends only on the initial and final temperatures and hence is the same as before.

$$S_2 - S_1 = 704 \text{ joules/deg}$$

Similarly, for the water at 323°K,

$$S_2 - S_1 = -603 \text{ joules/deg}$$

and the increase in entropy of the water is 101 joules/deg as before. However, the reservoirs are now part of the process, whereas in simple mixing only the two water-filled compartments are involved. Each reservoir gained or lost an amount of heat dQ; hence, each had an increase or decrease in entropy $dS = dQ/T$. The reservoirs as a group had a decrease in entropy of 101 joules/deg, thus making the total entropy of the system, consisting of two compartments plus all the reservoirs, a constant.

We can now restore the 2 kg of water at 323°K to the initial temperatures by reversing the process. If we take half the mass of water, 1 kg, and place it in one compartment of the apparatus, with the reservoir at $(323 - dT)°$ in the other, and remove the partition, the water will drop in temperature by an amount dT. Repeating this with the series of reservoirs down to 273°, we can restore this half of the water to its original state, and then go through a similar process with the other half. It should be obvious without calculation that the entropy of the water now decreases by 101 joules/deg and the entropy of the group of reservoirs increases by the same amount.

In comparing the reversible and irreversible experiments on the two 1-kg masses of water, we notice that in both cases energy is conserved, but that the effect on the entropy is different for each case. In the irreversible process, all the water ends up at the same temperature. No energy has been lost, but something else has: the opportunity of converting energy into useful work by connecting a machine to the two compartments 100° apart. Further, since the entropy increased and since $dU = T\,dS$, this entropy change must come by making an amount of energy dU unavailable for conversion into work. Since all real processes are irreversible, the fact that the entropy of the universe is increasing means that, although the energy of the universe is constant, it is becoming unavailable for conversion into work.

The water experiment provides another interpretation of the meaning of entropy: two compartments at two separate temperatures represent a system in a more highly organized state than the final mixture. Hence, we say that an increase in entropy is accompanied by an increase in *disorder*. A more precise definition of disorder will be given later in this chapter. For the moment, let us assume that we have defined disorder W in a quantitative way. Now we want the disorder to be proportional to the entropy. Further, if two systems have individual disorders W_1 and W_2, respectively, their combined disorder should be proportional to the product W_1W_2, since system 1 can have W_1 ways of being disordered for each way of system 2. The total entropy, however, of the two systems should be the sum of the individual entropies. A definition of disorder in terms of entropy which satisfies these requirements is

$$S = C \ln W \tag{2-20}$$

where C is a proportionality constant.

2-5 Enthalpy and the Helmholtz and Gibbs Functions

According to Eq. (2-3),

$$U = U(S, V, N_1, N_2) \tag{2-3}$$

the internal energy U is a function of S, V, and N. We say that the *natural variables* of U are S, V, and N. We would now like to find those functions which are natural variables of other sets of parameters; for example, the set T, V, and N. We note that $T = \partial U/\partial S)_{V,N}$ and we therefore have the problem of expressing U in the form

$$U = U(\partial U/\partial S, V, N) \tag{2-21}$$

This transformation is most easily accomplished by a mathematical device known as *the Legendre transform*. Consider a function $f(x,y)$ of x and y. Then

$$df = (\partial f/\partial x)\,dx + (\partial f/\partial y)\,dy \tag{2-22}$$

Now let

$$\partial f/\partial x \equiv f_x \qquad \partial f/\partial y \equiv f_y \tag{2-23}$$

and define g, the Legendre transform of f, by

$$g \equiv f - f_x x \tag{2-24}$$

Then

$$dg = df - f_x\,dx - x\,df_x$$

By Eq. (2-22), this reduces to

$$dg = f_y\,dy - x\,df_x \tag{2-25}$$

Comparing Eqs. (2-22) and (2-25) shows that if x and y are the natural variables of f, then f_x and y are the natural variables of g. Now

$$dg\,(f_x, y) = (\partial g/\partial f_x)\,df_x + (\partial g/\partial y)\,dy \tag{2-26}$$

Comparing Eqs. (2-25) and (2-26), we see that

$$\partial g/\partial f_x = -x, \qquad \partial g/\partial y = \partial f/\partial y \tag{2-27}$$

To apply this method to U and $\partial U/\partial S$, let

$$f = U, \qquad g = A, \qquad x = S, \qquad y = V$$

Then, by Eqs. (2-24) and (2-27),

$$A = U - (\partial U/\partial S)S$$

or

$$A = U - TS \tag{2-28}$$

where A is called the *Helmholtz function.*

The corresponding Legendre transform which expresses U in terms of S and $\partial U/\partial V = -P$ is called the *enthalpy H*

$$H = U + PV \tag{2-29}$$

Finally, we can express U in terms of both P and T by a simple extension of the above procedure, getting the *Gibbs function G*

$$G = U - TS + PV \tag{2-30}$$

Comparing the last three equations, we also have

$$G = A + PV = H - TS \tag{2-31}$$

Returning to Eq. (2-8), for a system with a single constituent

$$dU = TdS - P\,dV + K\,dN \tag{2-32}$$

and we see that the natural variables play the role of independent variables for the function $U(S, V, N)$.

PROBLEM 2-2 Show that the corresponding expressions for A, G, and H are

$$dA = -P\,dV - S\,dT + K\,dN \tag{2-33}$$

$$dG = -S\,dT + V\,dP + K\,dN \tag{2-34}$$

$$dH = T\,dS + V\,dP + K\,dN \tag{2-35}$$

PROBLEM 2-3

(a) By differentiation of $U(S, V, N)$ with respect to S and V, show that

$$\left.\frac{\partial T}{\partial V}\right)_S = -\left.\frac{\partial P}{\partial S}\right)_V \tag{2-36}$$

This is one of a set of equations known as *Maxwell's relations.*

(b) Find the other three equations in this set.

We would now like to consider the physical meaning of the three Legendre transforms A, G, and H. For constant N, by Eq. (2-33),

$$dA = -P\,dV - S\,dT$$

and for a constant-temperature process, this reduces to

$$dA = -P\,dV$$

By our sign convention, we see that dA represents the work done *on* the system during the process. For this reason, A is sometimes called the *free energy,* meaning the energy available for conversion into work. This name is very confusing, however, since both A and G have been referred to by chemists as free energy.

Next, by Eq. (2-34),

$$dG = -S\,dT + V\,dP$$

For a process during which P, T, and N remain constant, this shows that $dG = 0$, or G is constant. An example of such a process is the melting of ice or the boiling of water.

PROBLEM 2-4

(a) Differentiate $G(P, T, N)$ and compare with Eq. (2-34) to show that

$$V = \partial G/\partial P)_{T,N} \tag{2-37}$$

$$-S = \partial G/\partial T)_{P,N} \tag{2-38}$$

$$K = \partial G/\partial N)_{P,T} \tag{2-39}$$

These equations are analogous to Eqs. (2-5), (2-6), and (2-7)., In particular, Eq. (2-39) shows why K is also called the *Gibbs function per particle.*

(b) Go through the same manipulations with A to get

$$-S = \partial A/\partial T)_{V,N_1,N_2} \tag{2-40}$$

$$-P = \partial A/\partial V)_{T,N_1,N_2} \tag{2-41}$$

$$K_1 = \partial A/\partial N_1)_{V,T,N_2} \tag{2-42}$$

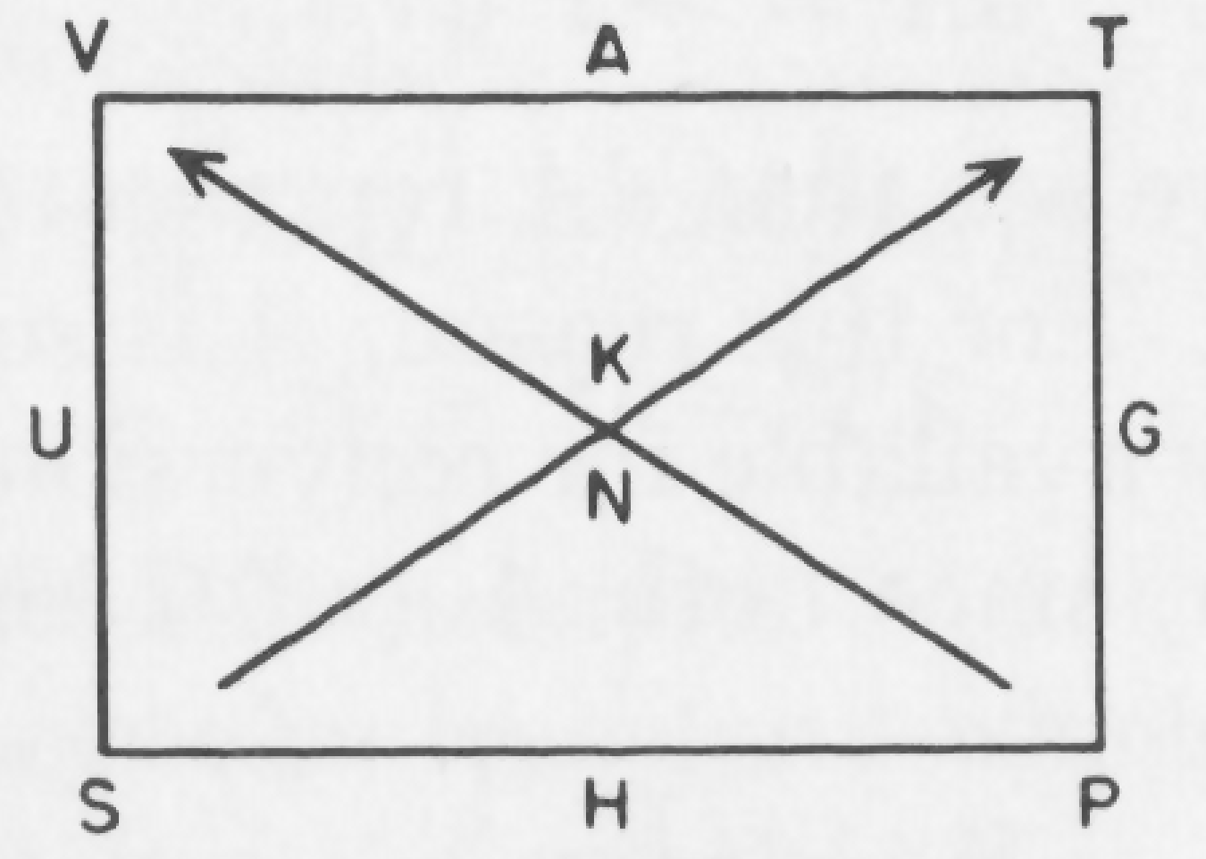
Hence, we see that K can also be called the *Helmholtz function per particle,* with respect to a different set of natural variables.

PROBLEM 2-5 Show that when N is constant

$$dH = dQ + V\,dP$$

Use this equation to explain why chemists refer to enthalpy as the "heat content" by applying it to a chemical reaction in an open container.

We will conclude our discussion of thermostatics with a useful mnemonic device which incorporates all the definitions and relations we have considered. In Fig. 2-2 we have placed U in the center of one side of a square, and the three Legendre transforms A, G, and H are at the centers of the other sides. The quantities S, V, T, and P are then placed on the corners so that U is between its natural variables S and V, and the other three functions lie between their own natural variables. Finally, we place K and N in the center, since N is a natural variable for all four: U, A, G, and H. The rules for using this $VATVUS$ diagram are as follows:



V = volume
T = temperature
P = pressure
U = internal energy
K = chemical potential

A = Helmholtz function
G = Gibbs function
S = entropy
H = enthalpy
N = number of particles

FIG. 2-2. The VATVUS diagram. (After H. B. Callen, *Thermodynamics*, John Wiley, New York, 1960, with permission.)

(a) When N is fixed

(1) Definitions of V, T, P, and S

To find P, for example, follow the arrow from P to V. Then

$$P = -\partial U/\partial V)_{S,N} = -\partial A/\partial V)_{T,N} \qquad \text{(2-6),(2-41)}$$

Similarly, to find T:

$$T = \partial U/\partial S)_{V,N} = \partial H/\partial S)_{P,N} \qquad \text{(2-5)}$$

etc.

(2) Definitions of A, U, H, and G

Start at the center of any side and go to the center of either adjacent side. Then move to the arrow which is parallel to the initial motion. For example,

$$A = U - TS = G - PV \qquad \text{(2-28), (2-31)}$$

$$G = H - TS = A + PV \qquad \text{(2-31)}$$

etc. The signs are determined by the arrows.

(3) Maxwell's Relations

To obtain Eq. (2-36), for example, start at P and move from S to V, then start at T and move from V to S.

$$-\frac{\partial P}{\partial S}\bigg)_V = \frac{\partial T}{\partial V}\bigg)_S$$

The sign is determined by whether the arrows are symmetric or antisymmetric with respect to the common side. As another example,

$$\frac{\partial S}{\partial V}\bigg)_T = \frac{\partial P}{\partial T}\bigg)_V$$

(4) The differential relations

Starting at the middle of any side, we pair up opposite corners:

$$dU = T\,dS - P\,dV \qquad \text{(2-32)}$$

The arrows determine the signs on each term.

Similarly,

$$dA = -S\,dT - P\,dV$$

etc.

(b) When N is variable

For the definitions of K,

$$K = \partial G/\partial N)_{P,T} = \partial U/\partial N)_{S,V} = \text{etc.},$$

the natural variables go with U, A, G, and H. The differential relations

$$dU = T\,dS - P\,dV + K\,dN$$

$$dA = -S\,dT - P\,dV + K\,dN$$

etc.

all include the term $K\,dN$.

2-6 The Statistical Distributions

We have seen that an electron in a crystal may have energies which are grouped into bands, but we have not been able to go any further than this.

We would like to know how many electrons have energies corresponding to each band. However, it will be realized immediately that it is impossible to measure or calculate the energies of the large number of electrons in a crystal; the best we can hope to do is calculate the probability that the electron will have some energy value E, where the dependence of this probability on E is called a *statistical distribution.*

To calculate a statistical distribution for a group of electrons in a solid, we shall adapt a method of Cowan (1957)[3] to the following simple problem: Figure 2-3 shows an egg carton containing n_1 square compartments, n_2 triangular compartments, n_3 round compartments, etc., or n_i of each kind. We also show a basket of N marbles. We now wish to calculate the number of different ways W in which the marbles can be arranged in the egg box, in accordance with the following three sets of rules (named after the scientists who essentially developed them):

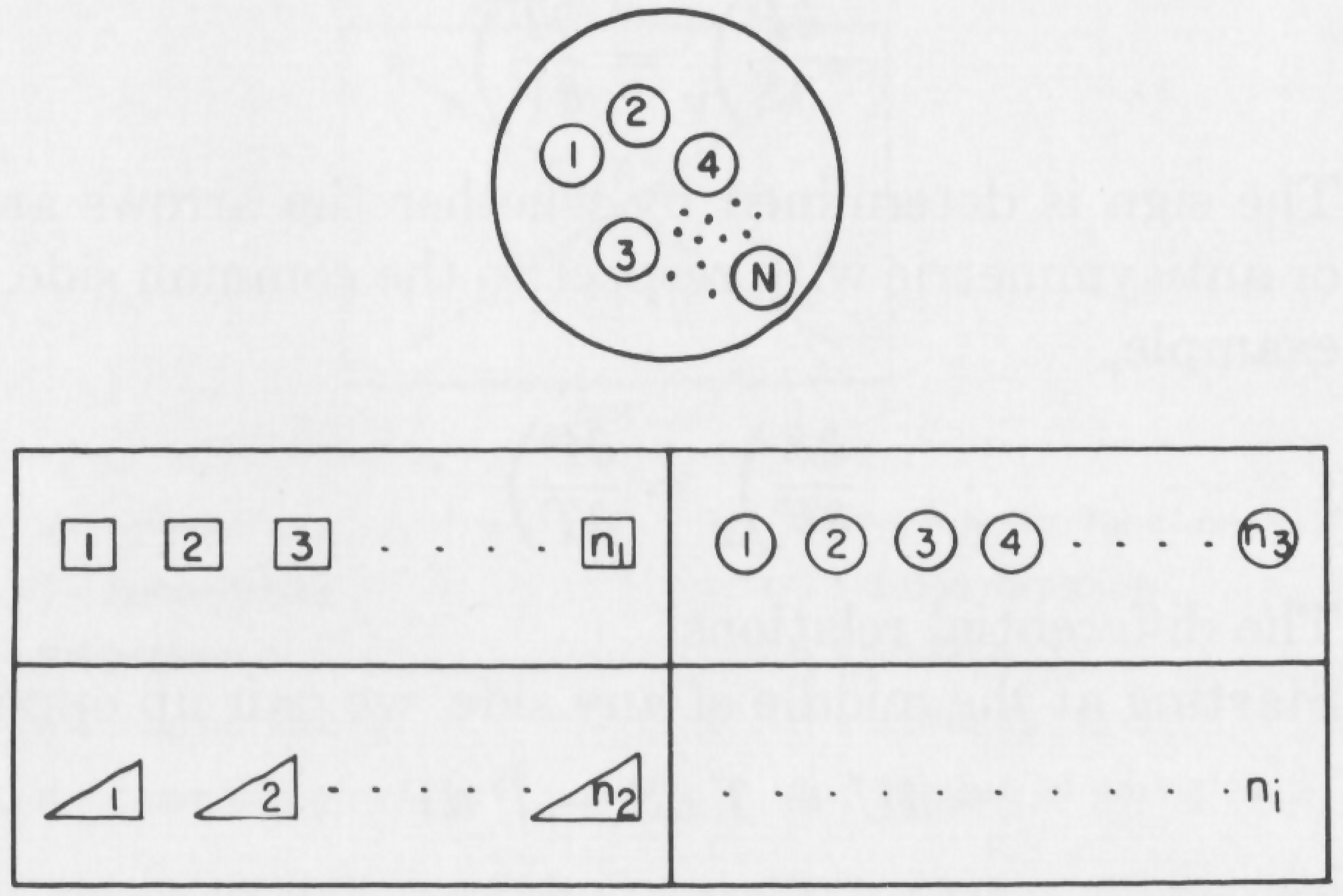


FIG. 2-3. A group of N distinguishable particles and a box containing n_1 square compartments, n_2 triangular compartments, and n_3 round compartments; that is, n_i of each type.

(1) *Maxwell-Boltzmann Distribution.* The marbles are made distinguishable by being numbered from 1 to N and there is no limit to the number permitted in each compartment.
(2) *Bose-Einstein Distribution.* The marbles are indistinguishable and there is no limit to the number permitted in each compartment.
(3) *Fermi-Dirac Distribution.* The marbles are indistinguishable and only one marble is permitted per compartment.

To each of the above rules we should add one condition that is often overlooked: the marbles are *non-interacting.* That is, when we apply these results to electrons, for example, their mutual repulsion is not taken into account.

To calculate W, we shall first subdivide the N marbles into groups, with N_1 in the first group, N_2 in the second group, and N_i in the ith group. The total number of ways in which the N marbles can be arranged among themselves is $N(N-1)\,(N-2)\ldots 1 = N!$, since the first marble can be any one of the N, the second one can be any one of the $(N-1)$ remaining, etc. Similarly, each of the groups of N_i marbles can be arranged in $N_i!$ ways; the total number of ways of arranging N_1 marbles in group number 1, N_2 in group number 2, and so on, is*

$$N_1!\,N_2!\ldots = \Pi N_i!$$

If W' denotes the number of ways of dividing the N marbles into groups, then

$$N! = W'\,\Pi N_i! \tag{2-43}$$

Now take the N_1 particles of group 1 and place them in the n_1 square compartments (with the restriction that we have chosen N_i to be less than n_i). There are n_1 ways of placing each of these particles into a compartment under the Maxwell-Boltzmann (MB) rules, or a total of $n_1^{N_1}$ ways of placing all N_1 and $\Pi n_i^{N_i}$ ways for all the groups. The number of possible Maxwell-Boltzmann distributions is the product of this quantity and W', or

$$W_{\mathrm{MB}} = N!\,\Pi(n_i^{N_i}/N_i!) \tag{2-44}$$

Since the electrons in a solid are considered to be indistinguishable, we can correct Eq. (2-44) for this situation if we divide W_{MB} by the number of ways $N!$ in which the N marbles can be rearranged among themselves, and the *corrected* Maxwell-Boltzmann distribution is

$$W_{\mathrm{MBC}} = \Pi(n_i^{N}/N!) \tag{2-45}$$

For the Bose-Einstein distribution, we will start by assuming the particles are distinguishable. Further, if we consider, for example, marbles number 1 and number 2 in the group of N_i marbles, the arrangement of putting number 1 in a compartment and then number 2 can be seen to be different from putting in number 2 and then number 1. Thus, there are n_1 ways of putting in the first marble, and $(n_1 + 1)$ ways for the second marble, since it may go into any of the empty $(n_1 - 1)$ compartments or into the occupied one in two ways. For the third marble, there are $(n_1 + 2)$ ways, etc., giving for all the groups $\Pi\, n_i\,(n_i + 1)(n_i + 2)\ldots(n_i + N_i - 1)$. We now divide by $N_i!$ to introduce the indistinguishability property

$$W_{\mathrm{BE}} = \Pi\,\frac{n_i(n_i + 1)(n_i + 2)\ldots(n_i + N_i - 1)}{N_i!} \tag{2-46}$$

---

*The symbol Π denotes the product of all the terms $N_i!$, where $i = 1, 2, \ldots$. It is analogous to the symbol Σ used to denote a sum.

For the Fermi-Dirac case, only one particle per cell is allowed. Then the first particle can go in any one of n_i cells, the second in any one of $(n_i - 1)$, etc., giving

$$W_{\text{FD}} = \Pi\, n_i(n_i - 1)\dots(n_i - N_i + 1)/N_i! \tag{2-47}$$

If we now define

$$\delta_{\text{MBC}} = 0, \qquad \delta_{\text{BE}} = -1, \qquad \delta_{\text{FD}} = +1$$

then Eqs. (2-45), (2-46), and (2-47) can be written as a single formula:

$$W = \Pi\, n_i(n_i - \delta)(n_i - 2\delta)\dots[n_i - (N_i - 1)\delta]/N_i! \tag{2-48}$$

PROBLEM 2-6

(a) The rules under which Eqs. (2-45) and (2-47) were derived are apparently the same. Why are the results different?

(b) Illustrate the cause of this difficulty by enumerating all the possibilities for the distribution of two marbles into three compartments and determine W_{MB}, W_{MBC}, W_{BE}, and W_{FD}.

We now identify the quantity W in the above equation with the disorder W of Eq. (2-20), so that the entropy can be expressed.

$$S = C\ln[\Pi\, n_i(n_i - \delta)\dots(n_i - N_i\delta + \delta)/N_i!] \tag{2-49}$$

We justify this definition by realizing that the more ways there are of rearranging a system, the more disorganized it is likely to be. We next wish to find the conditions under which S is a maximum, as required by the conditions stipulated in Section 2-2, and we will impose the following additional conditions or *constraints*:

(1) the total number N of marbles is fixed.

(2) the total energy E of the marbles is fixed.

Mathematically, these two constraints are written

$$\sum_j N_j = N \tag{2-50}$$

$$\sum N_jE_j = E \tag{2-51}$$

where E_j is the energy possessed by the group containing N_j marbles. (When we apply our results to electrons in solids, we shall use the terms *energy* and *internal energy* interchangeably, since we are primarily concerned with problems involving transfer of energy.)

The easiest way of maximizing S in Eq. (2-49) subject to the constraints [Eqs. (2-50) and (2-51)] is to use the *method of undetermined multipliers*.

Consider a function $f(x, y, z)$ of the three independent variables x, y, and z. This function has a maximum or minimum when

$$\partial f/\partial x = \partial f/\partial y = \partial f/\partial z = 0 \tag{2-52}$$

and we may solve Eq. (2-52) for the values x_0, y_0, z_0 at which $f(x, y, z)$ is an extremum. Suppose now that x, y, and z are connected by an additional relation, or constraint.

$$\varphi(x, y, z) = 0 \tag{2-53}$$

This new relation complicates the evaluation of the derivatives in Eq. (2-52). To get around this difficulty, we define a new function.

$$F(x, y, z) = f(x, y, z) + \lambda\varphi(x, y, z) \tag{2-54}$$

where λ is an unknown constant called *Lagrange's undetermined multiplier*. Then F will have an extremum when its derivatives are zero, or

$$\left.\begin{aligned} \partial f/\partial x + \lambda\partial\varphi/\partial x &= 0 \\ \partial f/\partial y + \lambda\partial\varphi/\partial y &= 0 \\ \partial f/\partial z + \lambda\partial\varphi/\partial z &= 0 \end{aligned}\right\} \tag{2-55}$$

These three equations plus Eq. (2-53) constitute four equations in the four unknowns x_0, y_0, z_0, and λ. If we have several constraints of the type of Eq. (2-53), then Eq. (2-55) will be of the form

$$\partial f/\partial x + \lambda_1\partial\varphi_1/\partial x + \lambda_2\partial\varphi_2/\partial x = 0 \tag{2-56}$$

PROBLEM 2-7 Use the method of undetermined multipliers to find the area of the largest rectangle that can be inscribed in a circle of unit radius.

For convenience, we will apply this method to ln W rather than to W. By Eq. (2-48),

$$\ln W = \sum_i \{\ln n_i + \ln(n_i - \delta) + \cdots + \ln[n_i - (N_i - 1)\delta]\}$$
$$-\sum_i (\ln 1 + \ln 2 + \cdots + \ln N_i)$$

Then $\partial(\ln W)/\partial N_i$, as N_i increases by unity, is

$$\partial(\ln W)/\partial N_i = \ln(n_i - N_i\delta) - \ln(N_i + 1) \tag{2-57}$$

By Eqs. (2-50) and (2-51), the two equations of constraint expressed in the form of Eq. (2-53) are

$$N - \sum N_j = 0$$
$$E - \sum N_jE_j = 0$$

so that
$$\frac{\partial}{\partial N_i}(N - \sum N_j) = -1$$

$$\frac{\partial}{\partial N_i}(E - \sum N_j E_j) = -E_i$$

Hence, Eq. (2-56) gives

$$\ln [(n_i - N_i\delta)/(N_i + 1)] - \lambda_1 - E_i\lambda_2 = 0 \tag{2-58}$$

and, if we neglect 1 in comparison with N_i,

$$\frac{N_i}{n_i} = \frac{1}{e^{\lambda_1 + E_i\lambda_2} + \delta} \tag{2-59}$$

This is the condition for the entropy to be a maximum that we were seeking.

2-7 The Perfect Gas

We now wish to evaluate two of the constants C and λ_2 of Eqs. (2-49) and (2-59). To do this, we will introduce the idea of a *perfect gas*, defined as a substance which obeys the combined Boyle-Charles Law:

$$PV = nRT \tag{2-60}$$

where n is the number of moles of the gas and R is called the universal gas constant. A mole is normally defined as that amount of gas whose mass in grams equals the molecular weight, but in the MKS system, we redefine it in terms of the kilogram. Then

$$R = 8.3 \times 10^3 \text{ joules/mole-deg}$$

If there are N molecules in the volume V, then

$$N = nL$$

where L, called *Avogrado's number*, has the value

$$L = 6.0 \times 10^{26} \text{ molecules/mole}$$

We now define *Boltzmann's constant* k as

$$\begin{aligned} k = R/L &= 1.4 \times 10^{-23} \text{ joule/deg} \\ &= 8.6 \times 10^{-5} \text{ eV/deg} \end{aligned} \tag{2-61}$$

and Eq. (2-60) becomes

$$PV = NkT \tag{2-62}$$

Return now to Eq. (1-87), which expresses the density of states in a box of volume $V = L^3$ as

$$dG = 4\pi Vp^2\, dp/h^3 \tag{1-87}$$

Let us now postulate the existence of a six-dimensional space called *phase space*, where three of the dimensions are the ordinary coordinates x, y, and z, and the other three are the momentum components p_x, p_y, and p_z. If an electron of mass m is moving in an arbitrary fashion, we can describe this motion by giving its position $\mathbf{r}$ and its momentum $\mathbf{p}$ at any instant or we can describe it equally well by specifying a single point in the six-dimensional phase space. An element of volume dV_P in phase space will be given by

$$dV_P = (L^3)(4\pi p^2\, dp)$$

so that the density of states is, from Eq. (1-87), simply

$$dG_P = \frac{1}{h^3} \tag{2-63}$$

PROBLEM 2-8 Show that Eq. (2-63) is dimensionally consistent.

To discuss the motion of a gas containing N molecules, we introduce a $6N$-dimensional phase space, and a single point in this space describes the motion of the entire system. Then, in a manner analogous to Section 1-10, the volume of a shell corresponding to the region between the energies E and $E + dE$ divided by the volume h^3 occupied by each point gives the total number of ways of distributing positions and momenta among these N molecules. This quantity, however, is what we have previously been calling W, so

$$\begin{aligned} W &= \frac{\int_E^{E+dE} dx_1\, dx_2 \ldots dx_{3N}\, dp_1\, dp_2 \ldots dp_{3N}}{h^{3N}} \\ &= (V/h^3)^N \int_E^{E+dE} dp_1\, dp_2 \ldots dp_{3N} \end{aligned} \tag{2-64}$$

To evaluate the integral in Eq. (2-64), we must take two steps:

(1) Define a *hypersphere* as the locus of all points in n-dimensional space such that

$$x_1^2 + x_2^2 + \cdots + x_n^2 = R^2 \tag{2-65}$$

where R is the radius. The surface area S of this sphere is then

$$S = KR^{n-1} \tag{2-66}$$

where K is a constant to be determined. Consider the integral

$$I = K \int_0^\infty e^{-R^2} R^{n-1}\, dR \tag{2-67}$$

PROBLEM 2-9 By letting $R^2 = t$, show that

$$I = \frac{K}{2}\left(\frac{n}{2} - 1\right)! \tag{2-68}$$

By Eq. (2-66), $KR^{n-1}\,dR = S\,dR$; this is a volume element in N-dimensional space.

PROBLEM 2-10 Replace $KR^{n-1}\,dR$ with $dx_1\,dx_2\,\ldots\,dx_n$ in Eq. (2-67) and show that

$$I = \pi^{n/2} \tag{2-69}$$

Combining Eqs. (2-67) and (2-69) yields

$$K = \frac{2\pi^{n/2}}{\left(\frac{n}{2} - 1\right)!} \tag{2-70}$$

Then the volume of the hypersphere is

$$V = K\int_0^R R^{n-1}\,dR = \frac{\pi^{n/2}R^n}{\left(\frac{n}{2}\right)!} \tag{2-71}$$

PROBLEM 2-11 Use Eq. (2-71) to find the "volume" of the figures for which $n = 1$, 2, and 3.

(2) We will next show that most of the volume of a hypersphere lies in a thin shell at the surface. The volume V_S of such a shell of thickness t is

$$\begin{aligned} V_S &= V(R) - V(R - t) \\ &= \frac{\pi^{n/2}}{(n/2)!}\left[R^n - (R - t)^n\right] \\ &= \frac{\pi^{n/2}R^n}{(n/2)!}\left[1 - \left(1 - \frac{t}{R}\right)^n\right] \end{aligned}$$

The term in brackets is approximately $(1 - e^{-nt/R})$; for large values of n, we will have $nt \ll R$, so that V_S is approximately equal to V.

PROBLEM 2-12 Verify this approximation by using the Taylor series for e^x given in Appendix A1.

Applying the above results, we integrate Eq. (2-64) over the entire $3N$-dimensional hypersphere of radius p^2, obtaining

$$W = \frac{(V/h^3)^N\left[\pi^{3N/2}\,(2mE)^{3N/2}\right]}{(3N/2)!} \tag{2-72}$$

PROBLEM 2-13

(a) Using Eqs. (2-20) and (2-72), and the *VATVUS* diagram of Fig. 2-2, show that

$$\frac{1}{T} = \left.\frac{\partial S}{\partial U}\right)_{V,N} = \frac{3NC}{2E} \tag{2-73}$$

What does this result mean?

(b) Using the identity from differential calculus

$$\left.\frac{\partial x}{\partial y}\right)_z \left.\frac{\partial y}{\partial z}\right)_x \left.\frac{\partial z}{\partial x}\right)_y = -1 \tag{2-74}$$

and the formulas mentioned in part (a), show that

$$\frac{P}{T} = \left.\frac{\partial S}{\partial V}\right)_{U,N} = \frac{NC}{V} \tag{2-75}$$

Comparing Eq. (2-75) with Eq. (2-62) shows that

$$C = k$$

and

$$S = k \ln W \tag{2-76}$$

This makes Eqs. (2-73) and (2-75) become

$$E = \tfrac{3}{2}NkT \tag{2-77}$$

$$PV = NkT \tag{2-78}$$

At the beginning of this section we stated that Eq. (2-78) served to define a perfect gas. We see now that this equation is a necessary part of this definition, but is not alone sufficient, since Eq. (2-77) must also be included. Further, in our derivation of these equations, we have considered only the linear momentum p, and have ignored any angular motion of the gas molecules. This procedure is valid only for a monatomic gas, and the energy E is restricted to translational energy.

Having identified C as the Boltzmann constant k, we now go on to consider λ_2. Since W is a function of the N_j, the differential of W is

$$d(\ln W) = \sum_j \left(\frac{\partial}{\partial N_j}\ln W\right) dN_j \tag{2-79}$$

By Eqs. (2-57) and (2-58),

$$\frac{\partial}{\partial N_i} \ln W = \lambda_1 + E_i\lambda_2 \tag{2-80}$$

Hence

$$dS = k\, d(\ln W)$$

$$= k \sum_j (\lambda_1 + E_j\lambda_2)\, dN_j \tag{2-81}$$

By Eq. (2-50),

$$\sum_j dN_j = 0$$

and Eq. (2-81) becomes

$$dS = k\lambda_2 \sum_j E_j\, dN_j$$

or, by Eq. (2-51),

$$dS = k\lambda_2\, dE \tag{2-82}$$

Since we are holding both N and V constant, Eq. (2-82) is equivalent to Eq. (2-5), and

$$\lambda_2 = 1/kT \tag{2-83}$$

2-8 The Fermi Level and the Electrochemical Potential

We shall deal with the remaining undetermined multiplier λ_1 of Eq. (2-59) in a way that differs from that used for λ_2 and C. Let us define an energy E_F, called the *Fermi energy* or *Fermi level,* by

$$\lambda_1 = -E_F/kT \tag{2-84}$$

Using this definition and Eq. (2-83), Eq. (2-59) becomes, for the Fermi-Dirac distribution,

$$\frac{N_i}{n_i} = \frac{1}{e^{(E_i - E_F)/kT} + 1} \tag{2-85}$$

The right-hand side of Eq. (2-85) is called the *Fermi function* $f(E_i)$; the purpose of this section will be to establish a physical interpretation of E_F and $f(E_i)$.

Equation (2-85) was derived by considering the problem of distributing N_i marbles with energy E_i into n_i compartments. As a result of experiments to be described in Chapters 7 and 8, it has been verified that electrons in impure semiconductors and in metals obey the Fermi-Dirac rules. To apply Eq. (2-85) to electrons, then, we identify n_i as the number of energy levels lying in a small region dE about the value E, which we have shown to be

$$dG = 4\pi p^2\, dp\, V/h^3 \tag{1-87}$$

However, this formula should be modified to include spin. Considering spin, the exclusion principle permits two electrons per set (n_x, n_y, n_z) of quantum numbers for the electron in the box just as it did for the set (n, l, m_l) applying to electrons in an atom, as described in Section 1-12. Hence, Eq. (1-87) becomes

$$dG = 8\pi p^2\, dp\, V/h^3 \tag{2-86}$$

Next, we identify N_i as dN, the number of electrons whose energy lies in a region dE about E, where we have dropped the subscript i on E. Then Eq. (2-85) becomes

$$dN = \frac{8\pi\, p^2\, dp\, V/h^3}{e^{(E-E_F)/kT} + 1} = f(E)\, dG \tag{2-87}$$

It now seems very reasonable to interpret $f(E)$ as the probability that an electron has an energy E, since the product of this probability and the number of available states dG should give dN, the total number of electrons whose energies lie in this range.

PROBLEM 2-14

(a) Show that for $T = 0°\text{K}$

$$f = 1 \text{ when } E < E_F$$

$$f = 0 \text{ when } E > E_F$$

(b) Show that for any T, $f = \frac{1}{2}$ when $E = E_F$.

This problem enables us to plot $f(E)$ as a function of E at $T = 0°\text{K}$, as shown in Fig. 2-4. By part (b), we see that the Fermi level E_F is the energy level which has a 50 per cent probability of being occupied.

PROBLEM 2-15

(a) Explain qualitatively why the curve for $T = 300°\text{K}$ in Fig. 2-4 differs from the curve for $T = 0°\text{K}$.

(b) Prove that the curve for $T = 300°\text{K}$ is symmetric about the line $f = \frac{1}{2}$.

We can get an insight into the meaning of the Fermi level at $T = 0°\text{K}$ by realizing that, at this temperature, the electrons should have their lowest possible energy. Because electrons in a solid are subject to the exclusion principle, and hence to the Fermi-Dirac rules, and since these rules permit only one electron per energy level (neglecting spin for the moment), all levels up to a certain value of energy will be filled, and those beyond will be empty. This value of energy is denoted E_{F0}, the *Fermi level*

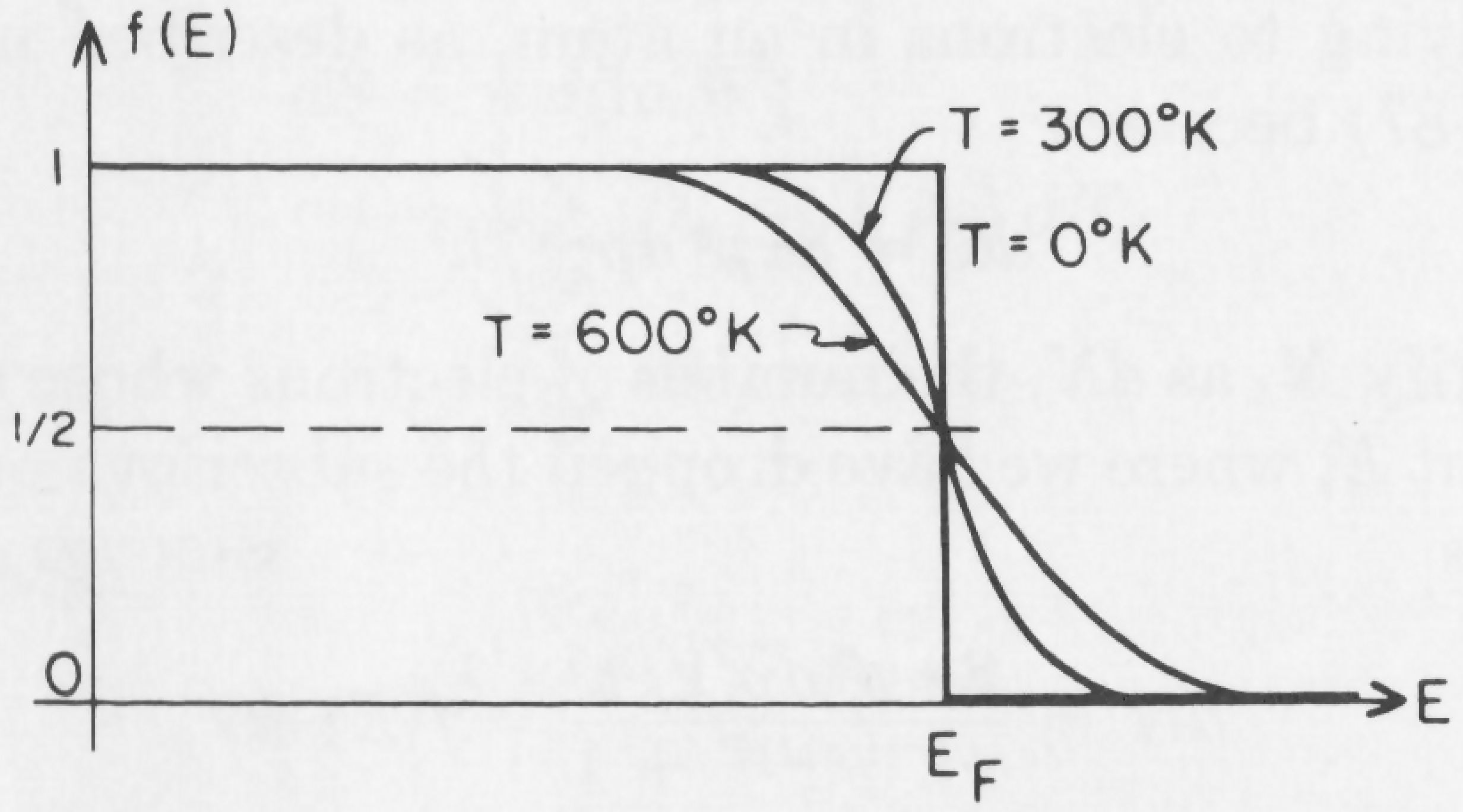


FIG. 2-4. The Fermi function and its dependence on temperature. (From E. Spenke, *Electronic Semiconductors*, McGraw-Hill, New York, 1958, with permission.)

at absolute zero. To estimate the value of E_{F0}, we calculate the radius R_m of the sphere in Section 1-10 corresponding to these occupied states. The volume of the positive octant is

$$\tfrac{1}{8}\,(4\pi R_m^3/3) = \pi R_m^3/6$$

By Eqs. (1-84) and (1-85),

$$E_m = (h^2/8mL^2)\,R_m^2 = E_F$$

Since all points in the octant represent occupied levels, the number of levels is twice the number N of electrons, including spin, and

$$\frac{N}{2} = \frac{\pi}{6}\left(\frac{8mL^2E_F}{h^2}\right)^{3/2}$$

or

$$E_{F0} = (3n/\pi)^{2/3}\,(h^2/8m) \tag{2-88}$$

where

$$n = N/L^3$$

is the number of electrons per unit volume.

For silver, $n = 5.9 \times 10^{28}/m^3$, and Eq. (2-83) gives

$$\begin{aligned} E_{F0} &= (3 \times 5.9 \times 10^{28}/\pi)(6.6 \times 10^{-34})^2/8 \times 9 \times 10^{-31} \\ &= 9.0 \times 10^{-19}\,\text{joule} = 5.6\ \text{eV} \end{aligned}$$

and we have assumed that m_e is approximately equal to m.

The Fermi energy E_F has another interpretation which is directly connected to the thermostatic functions. By Eq. (2-28),

$$A = E - TS \tag{2-28}$$

For constant temperature and volume

$$dA = dE - T\,dS$$

By Eqs. (2-81), (2-83), and (2-84),

$$\begin{aligned} dA &= dE - Tk \sum [(-E_F/kT) + (E_j/kT)]\,dN_j \\ &= dE + E_F \sum_i dN_j - \sum E_j\,dN_j \end{aligned}$$

If we remove the restriction that N is a constant, the second term on the right does not vanish, although the third term does. Hence,

$$dA = E_F\,dN \tag{2-89}$$

For constant T and V, this is equivalent to

$$\partial A/\partial N)_{T,V} = E_F \tag{2-90}$$

Comparing this with Eq. (2-42), we identify the Fermi energy of statistical physics with the chemical potential of thermostatics.

In developing the subjects of thermostatics and statistical physics, it should again be emphasized that we have been dealing with *uncharged* particles. Even though we have talked about electrons, we have not made use of the fact that they each carry a charge $-e$. We now take this property into account in Eq. (2-3) by incorporating charge q into E as an additional natural variable.

$$E = E(S, V, N, q) \tag{2-91}$$

Then, as in Section 2-3,

$$dE = T\,dS - P\,dV + K\,dN + \mathcal{V}\,dq \tag{2-92}$$

where

$$\mathcal{V} \equiv \partial E/\partial q)_{S,V,N} \equiv \textit{electrostatic potential} \tag{2-93}$$

which is a more general form of Eq. (1-15). We can add this term to the definitions of A, G, and H in the same manner in which we added the $K\,dN$ terms in Problem 2-2. For example, Eq. (2-34), for a system containing one chemical type, will be

$$dG = V\,dP - S\,dT + K\,dN + \mathcal{V}\,dq \tag{2-94}$$

We now define the *electrochemical potential* $\bar{\mu}$ ("μ-bar") as

$$\bar{\mu} = K - e\mathcal{V} \tag{2-95}$$

This definition states that the electrochemical potential $\bar{\mu}$ is the sum of the chemical potential K and the electric potential $\mathcal{V}$ (multiplied by $-e$ to express it in energy units and to incorporate the fact that the charge on the electron is negative). In other words, the total work involved in moving an electron should be the sum of the chemical and the electrical work. Substituting K from Eq. (2-95) into Eq. (2-94) shows that

$$dG = V\,dP - S\,dT + \bar{\mu}\,dN \tag{2-96}$$

if we use the fact that

$$dq = -e\,dN$$

That is, the total charge dq is the product of the charge $-e$ on each electron times the number dN of electrons. Now Eq. (2-96) could also be obtained from Eq. (2-34) by replacing K with $\bar{\mu}$. In general, all four previous equations for uncharged particles containing K can be made valid for charged particles by substituting $\bar{\mu}$. An important consequence of this fact is that Eq. (2-42) now becomes

$$\bar{\mu} = \partial A/\partial N)_{T,V} \tag{2-97}$$

and by Eq. (2-90),

$$\bar{\mu} = E_F \tag{2-98}$$

Hence, under equilibrium conditions, the electrochemical potential is identical with the Fermi level. We shall discuss the non-equilibrium situation in a later chapter in connection with the properties of semiconductor junctions.

PROBLEM 2-16

(a) Show that at room temperature

$$kT = 0.026 \text{ eV}$$

(b) For what range of values of E can $f_{FD}(E)$ be replaced by the simpler $f_{MB}(E)$?

(c) Using the results of Problem (2-13), find the average energy in eV of a molecule in a gas at 50,000°K. Compare this to the maximum energy of an electron in a metal at 0°K.

A physical interpretation of the electrochemical potential $\bar{\mu}$ can be obtained by considering an extension of Ohm's law.

$$J = \sigma\mathcal{E} \tag{1-21}$$

The inverse of Eq. (1-16) is

$$\mathcal{E} = -\frac{d\mathcal{V}}{dx} \tag{2-99}$$

so that

$$J = -\sigma\frac{d\mathcal{V}}{dx} \tag{2-100}$$

or

$$eJ = -e\sigma\frac{d\mathcal{V}}{dx} \tag{2-101}$$

By definition [Eq. (2-95)], if we assume for the moment that $K = 0$, Eq. (2-101) becomes

$$eJ = \sigma\frac{d\bar{\mu}}{dx} \tag{2-102}$$

In establishing Ohm's law, we regarded electrons as particles each carrying a charge $-e$, but having no chemical properties (that is, no mass). If we now consider both the electrical and chemical aspects of the electrons, then it seems reasonable to assume that Eq. (2-102) holds even when $K \neq 0$. Substituting the definition [Eq. (2-95)] into Eq. (2-102)

$$eJ = \sigma\left(\frac{dK}{dx} - e\frac{d\mathcal{V}}{d}\right)$$

or

$$J = \left(\frac{\sigma}{e}\right)\frac{dK}{dx} + \sigma\mathcal{E} \tag{2-103}$$

This relation is a more generalized form of Ohm's law, applicable when both chemical and electrical forces are present. Writing Eq. (2-102) as

$$J = \left(\frac{\sigma}{e}\right)\frac{d\bar{\mu}}{dx} \tag{2-104}$$

and comparing this with Eq. (2-101), we see that the current created by the motion of pure charges is due to a gradient in the electric potential, but that the current for particles with which we associate both charge and mass is due to the electrochemical potential gradient.

In a later chapter, we shall discuss a further generalization which takes into account temperature gradients as well.

REFERENCES

1. Zemansky, M. W., *Heat and Thermodynamics.* New York: McGraw-Hill Book Co., 1937.

2. Sears, F. W., *Thermodynamics, Kinetic Theory, and Statistical Mechanics.* Cambridge: Addison-Wesley, 1950.

3. Cowan, R. D., *Am. J. Phys.*, **25,** 463 (1957).

GENERAL REFERENCES

Callen, H. B., *Thermodynamics.* New York: John Wiley & Sons, Inc., 1960.
Kittel, C., *Elementary Statistical Physics.* New York: John Wiley & Sons, Inc., 1958.

In this chapter, we have restricted our discussion to the very minimum necessary for an understanding of semiconductors. These two references are recommended for those readers who would like to have a better understanding of thermodynamics and statistical mechanics.

3

ENERGY BANDS AND CARRIER STATISTICS

3-1 Conductors, Insulators, and Semiconductors

In Chapter 1, we discussed the dual nature of the electron and showed that its motion or propagation is governed by the Schroedinger equation rather than by the classical wave equation. In solving this equation for simple cases like the square well, we found that the total energy was *quantized* — that is, restricted to certain definite values. For the periodic potential energy of the one-dimensional crystal, the permitted energy values were spread out into bands. We now wish to consider in more detail the nature of these bands and their connection with the electrical properties of crystalline solids.

Let us return to Eq. (1-111) and observe that the right-hand side, cos kL, is a periodic function of k. That is, when k is replaced by $k + 2\pi n/L$, where n is an integer, then

$$\cos (k + 2\pi n/L)L = \cos kL$$

This means that the energy bands shown in Fig. 1-21 can be generated by values of k restricted to an interval of width $2\pi/L$. Choosing this interval as

$$-\pi/L \leqq k \leqq \pi/L \tag{3-1}$$

we can replot Fig. 1-21b as shown in Fig. 3-1a.

The one-dimensional crystal of Fig. 1-20 was taken to be periodic and infinite. By considering this infinite crystal, we could get around the difficulty that the periodic properties are destroyed at the ends. We can also avoid this difficulty for a finite crystal with N atoms by postulating *cyclic boundary conditions*. Let us take a one-dimensional crystal with N atoms and bend it around in a circle so that the periodic property holds everywhere (Fig. 3-1b). The length of this crystal is NL and the boundary condition, when applied to the function $\psi(x)$ of Eq. (1-102), gives

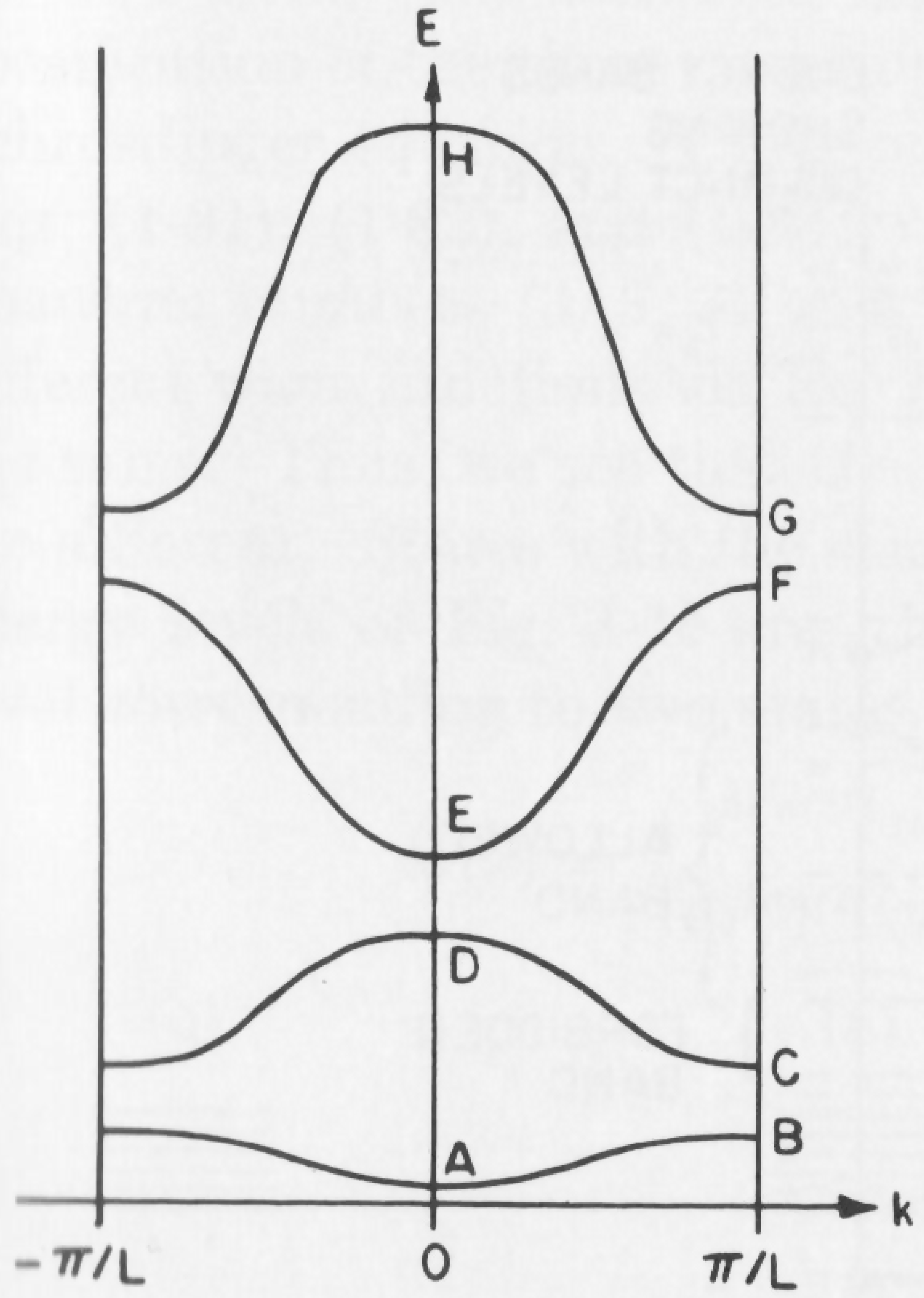


FIG. 3-1. (a) An alternate form of the energy bands of Figure 1-21. (From A. J. Dekker, *Solid State Physics*, Prentice-Hall, Englewood Cliffs, N.J., 1957, with permission.)

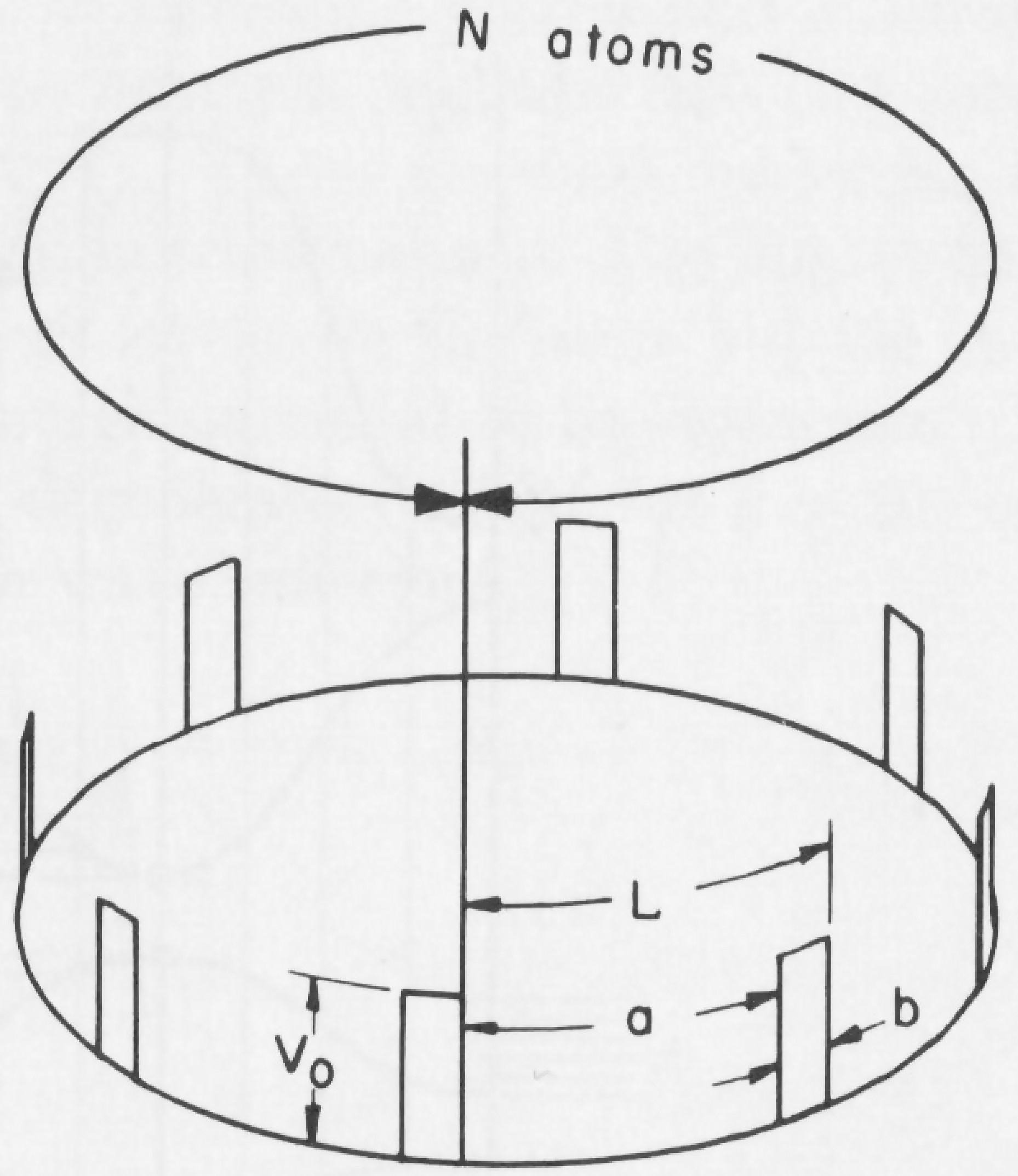


FIG. 3-1. (b) The periodic lattice of Figure 1-20 rearranged to form a circle.

$$\psi(x + L) = \psi(x)$$

or

$$e^{ik(x+NL)}\,u(x + NL) = e^{ikx}u(x) \tag{3-2}$$

Since $u(x)$ is required to have a period L, Eq. (3-2) reduces to

$$e^{ikNL} = 1$$

or $kNL = 2\pi n$ and

$$k = 2\pi n/NL \quad \text{where } n = 0, \pm 1, \pm 2, \ldots \tag{3-3}$$

Equation (3-3) is based on the fact that $e^{2\pi i} = 1$, as can be seen from Appendix A-1. Combining Eqs. (3-1) and (3-3) shows that the integer n lies in the range of values given by

$$-N/2 \leqq n \leqq N/2 \tag{3-4}$$

When N is a large number, Eq. (3-4) shows that there are N possible values of n, and hence of k. Figure 3-1c indicates that there is an energy value for every pair of values $\pm 2\pi n/NL$ of k in the interval $-\pi/\mathrm{L}$ to π/L, giving $N/2$ energy levels in each band. The consequence of the periodic boundary conditions has thus been to demonstrate that, in an ideal crystal, even in the allowed bands, only a finite number of discretely-spaced energies are available to the electrons.

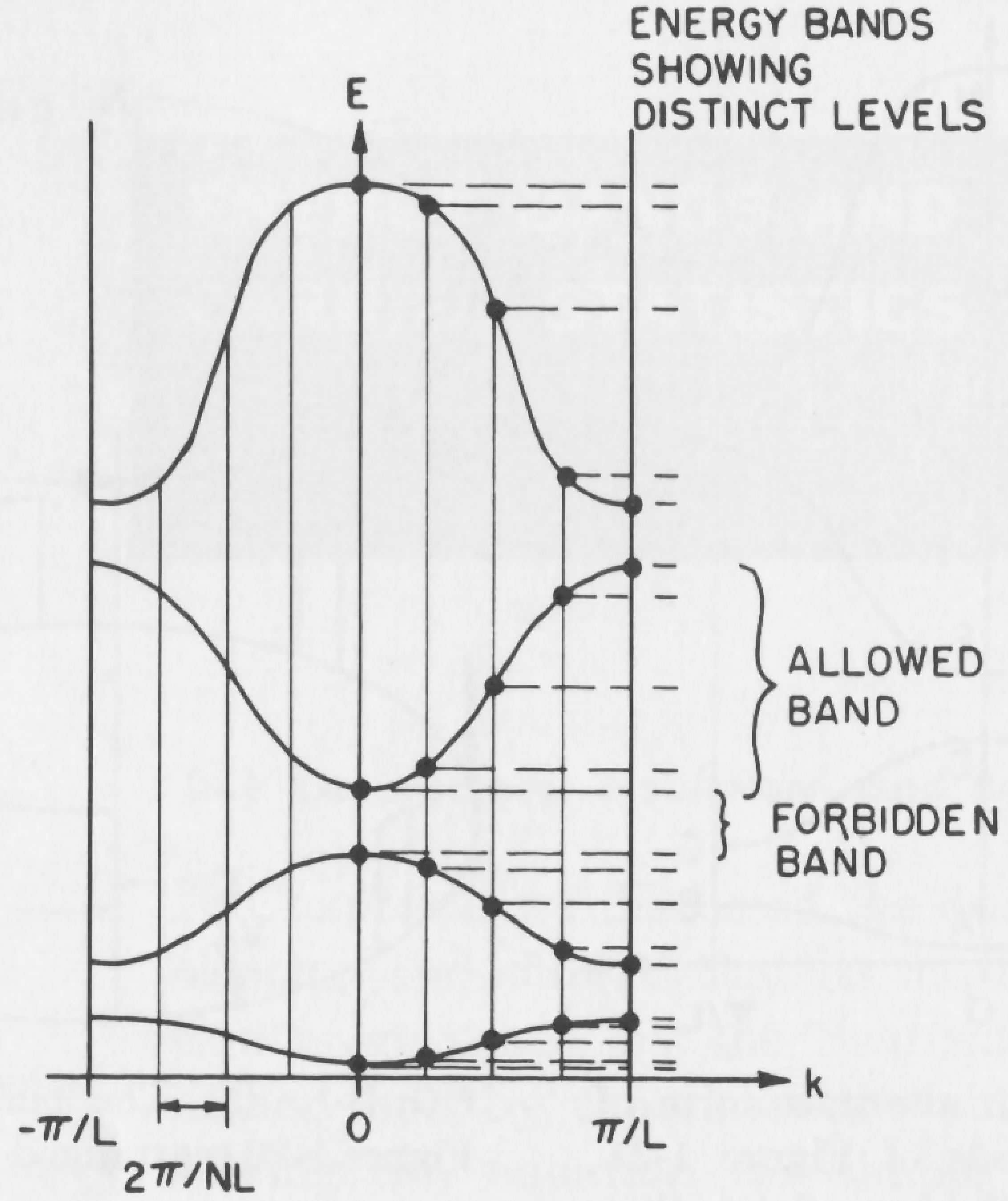


FIG. 3-1. (c) The discrete levels permitted by periodic boundary conditions. (From W. Shockley, *Electrons and Holes in Semiconductors*, D. Van Nostrand, New York, 1950, with permission.)

PROBLEM 3-1

(a) Estimate the width in eV of the widest band in Fig. 3-1 for a hypothetical one-dimensional hydrogen crystal.

(b) Calculate the spacing in eV of the levels in this band.

This idea of the discrete structure of energy bands can also be obtained in a less mathematical way by considering the multielectron atoms covered in Section 1-12. It was pointed out there that the mutual repulsion of the electrons in these atoms shifts the original hydrogen levels. Bringing a group of these identical atoms together to form a periodic crystal should result in allowed and forbidden bands; the allowed bands will consist of a large number of distinct levels because of a further shift in the atomic energies due to interaction of electrons from neighboring atoms.

Before considering the effect of the band structure on the electrical properties of crystals, it will be well to emphasize the distinction between the two terms *state* and *level*. In Section 1-10, we defined a state as the specification of the three quantum numbers n_x, n_y, n_z in the solution to the Schroedinger equation. These numbers determine both ψ and E through Eqs. (1-81), (1-82), and (1-84). As an example, consider the two sets of quantum numbers (3, 3, 3) and (5, 1, 1). These numbers determine two different wave functions via Eq. (1-81), but the value of E in Eq. (1-84) is the same. Thus, we see that the energy levels are the same but the states are different. States with the same energies are said to be *degenerate*. The energy levels of Fig. 3-1c are, then, degenerate (except for $k = 0$), each level corresponding to two states with wave functions.

$$\psi(x) = \begin{Bmatrix} e^{2\pi n/NL}u(x) \\ e^{-2\pi n/NL}u(x) \end{Bmatrix} \qquad n = 0, 1, 2, \ldots$$

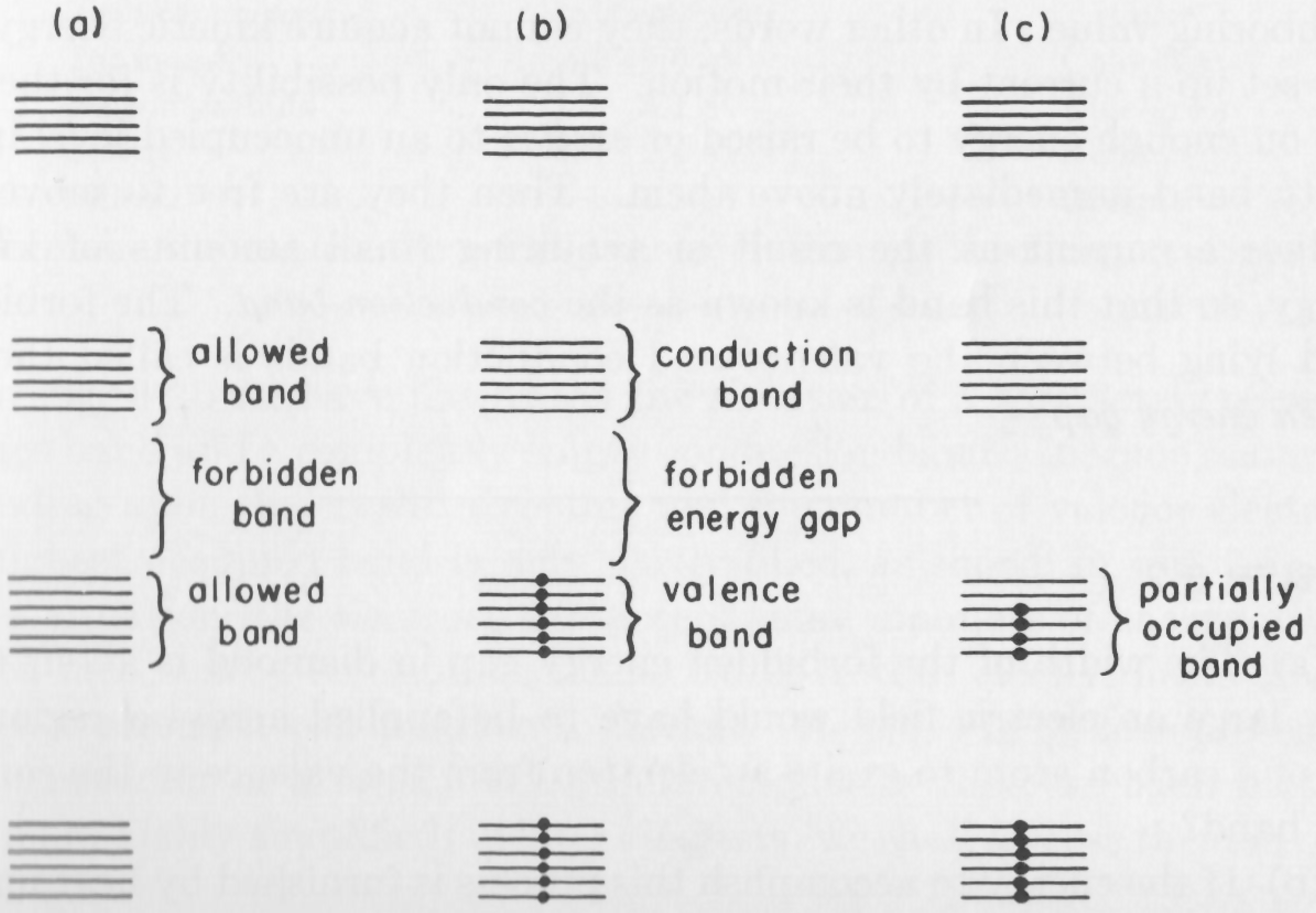


FIG. 3-2. (a) A schematic representation of the allowed and forbidden bands in crystals. (b) The band arrangement in an insulator. (c) Metallic energy bands.

The integer n is a quantum number related to k through Eq. (3-3), just as k_x is related to n_x through Eq. (1-82).

Now consider the energy bands in a real crystal. In Fig. 3-2a we have reproduced the arrangement of levels taken from Fig. 3-1c. If we cool the crystal to very low temperatures, the electrons will have their energies reduced to the lowest values possible. If we neglect spin and degeneracy, each level corresponds to a distinct state. The exclusion principle then results in the situation shown in Fig. 3-2b with one electron per level up to a certain maximum and all above that unoccupied. We have deliberately taken the uppermost occupied level in this diagram as the top of the second band. In the separated atoms from which this crystal was formed, the electrons are in the ground state at 0°K. Although higher levels are available, they are not occupied, and we should expect a similar situation to prevail in the crystal. That is, the upper empty band corresponds to the excited states of electrons, and the highest occupied band corresponds to the ground state of the outermost, or *valence*, electrons in the atom. For this reason, the upper occupied band is called the *valence band*, and a crystal whose energy bands look like Fig. 3-2b is generally a non-conductor of electricity, or an *insulator*. To see why this is so, let us apply a voltage across the ends of the crystal (still holding the temperature at 0°K). Since all the energy levels in the valence band are occupied, the electrons in this band are not permitted by the exclusion principle to take on small amounts of energy from the electric field and change their own energy to some neighboring value. In other words, they cannot acquire kinetic energy and thus set up a current by their motion. The only possibility is for them to take on enough energy to be raised or *excited* to an unoccupied level in the empty band immediately above them. Then they are free to move and produce a current as the result of acquiring small amounts of kinetic energy, so that this band is known as the *conduction band.* The forbidden band lying between the valence and conduction bands is called the *forbidden energy gap.*

PROBLEM 3-2

(a) The width of the forbidden energy gap in diamond is about 6 eV. How large an electric field would have to be applied across a region the size of a carbon atom to excite an electron from the valence to the conduction band?

(b) If the energy to accomplish this process is furnished by heat instead of by an electric field, to what temperature would the crystal have to be raised? [Hint: To find the approximate energy, use Eq. (2-77).]

In the above problem, it is seen that diamond, for example, requires tremendous electric fields or a high temperature to excite electrons into the conduction band, and hence is an insulator. On the other hand, gray tin is found experimentally to have a gap width of about 0.1 eV. As the crystal is brought up to room temperature, the Fermi-Dirac distribution makes the change indicated in Fig. 2-4. The energies of the electrons are spread over a range about 2 kT wide, which is about 0.05 eV, according to Problem 2-16. Some of the electrons will then have enough energy to be raised into the conduction band. By the application of heat, we have converted an insulator into a conductor; such materials are known as *intrinsic semiconductors.* The process of raising an electron into the conduction band by applying heat is known as *thermal excitation;* the analogous process using light as the source of energy is called *optical excitation.*

PROBLEM 3-3

(a) Show that the energy E in eV associated with light of wave length λ in microns is given by

$$E = 1.24 \text{ eV/micron} \tag{3-5}$$

(b) Table P3-1 lists four semiconductors, their approximate energy gap, and their color. Verify the formula of Eq. (3-5) in a qualitative way and explain the table. To what type of radiation is silicon transparent?

TABLE P3-1

Semiconductor	*Energy gap*	*Color*
Diamond	5.6 eV	Colorless
Silicon carbide	3.1 eV	Blue-green
Sulfur	2.4 eV	Yellow
Silicon	1.1 eV	Opaque

In Fig. 3-2b we have illustrated the situation of a completely occupied valence band and a completely empty conduction band. In some materials, depending upon the crystal structure and the number of valence electrons, the highest occupied band is only partly filled, as shown in Fig. 3-2c. In such a situation, the electrons can accept small amounts of energy and be excited into nearby levels in the same band, so that electric fields of even very low strength will produce a current. Hence, Fig. 3-2c represents a good conductor or a *metal.* We should emphasize that the band pictures shown are highly simplified; in later chapters, we shall discuss them in more detail.

In addition to intrinsic semiconductors, there is another class called *impurity semiconductors.* In Table 3-1 we have reproduced a section of the periodic table; the elements of column IV which are shown — carbon, silicon, germanium, and gray tin — are semiconductors. Each of these elements has four electrons in its outer, or valence, shell. These elements form *covalent* crystals; each atom is joined to its four neighbors by sharing an electron-pair, as shown in Fig. 3-3.

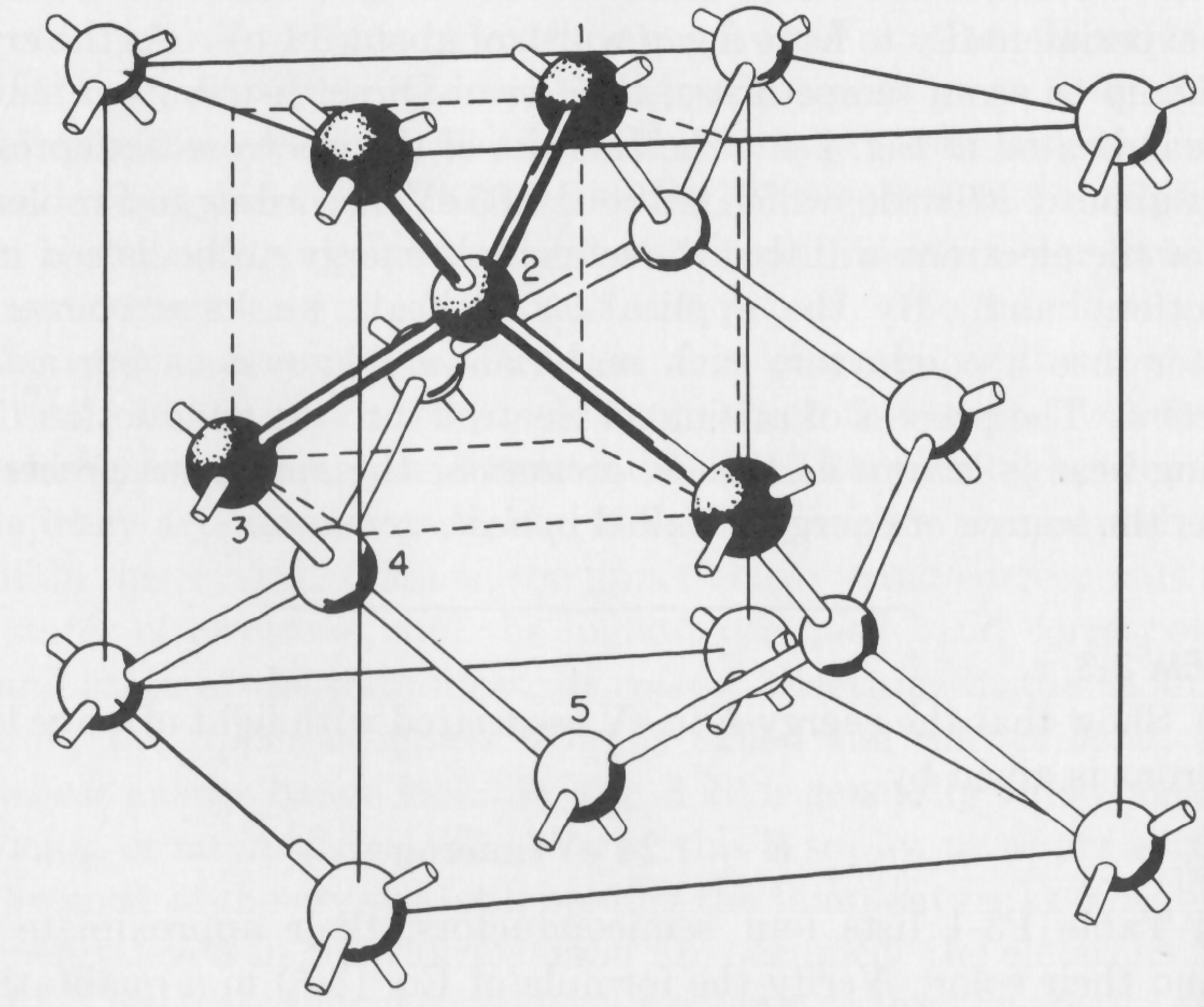


FIG. 3-3. The diamond lattice; the numerals indicate a *screw axis*, which is discussed in Chapter V. (From V. Heine, *Group Theory in Quantum Mechanics*, Pergamon, New York, 1960, with permission.)

TABLE 3-1

IIIA	IVA	VA	VIA	VIIA	Inert gases
5 B	6 C	7 N	8 O	9 F	10 Ne
13 Al	14 Si	15 P	16 S	17 Cl	18 A
31 Ga	32 Ge	33 As	34 Se	35 Br	36 Kr
49 In	50 Sn	51 Sb	52 Te	53 I	54 Xe
81 Tl	82 Pb	83 Bi	84 Po	85 At	86 Rn

Let us take one of these crystals, remove one of the column IV atoms, and replace it with an atom from column V, say phosphorus. Four of the five valence electrons of phosphorus form covalent bonds with the four

surrounding atoms, and the fifth electron is free. This electron is, therefore, called an *excess electron*. Intuitively, we would expect that if an electric field is applied to this crystal, the electron will move and carry current. This is another mechanism by which an insulator may be converted into a conductor, and these materials are called *impurity semiconductors*. The impurity is called a *donor*, since it contributes conduction electrons. It should be noted that the same crystals can exhibit both intrinsic and impurity conduction. At low temperatures, most of the conduction electrons come from the donor impurities. As the temperature is increased, the number of electrons excited directly from the valence band to the conduction band increases, and the intrinsic conductivity is the dominant process.

It is interesting to compare the conductivity σ of metals, insulators, and semiconductors. Typical values are shown in Table 3-2, which demonstrates the wide range of this property.

TABLE 3-2

Material	σ
Metal (copper)	1×10^{8} ohm^{-1} m^{-1}
Semiconductor (germanium)	20 ohm^{-1} m^{-1}
Insulator (sulfur)	1×10^{-15} ohm^{-1} m^{-1}

We should point out that the diagrams of Fig. 3-2 are not graphs, since only the variable E is involved. They simply represent relative positions of energy levels on a vertical scale. We ask how these energy levels depend upon the position of the conduction electron in the crystal. Let us recall that the solutions to the Schroedinger equation of Chapter 1 showed that the electron has a small, but finite, probability of being found outside the wells which make up the periodic potential. Hence, the plot of energy versus position should be continuous throughout the crystal and should appear as shown in Fig. 3-4.

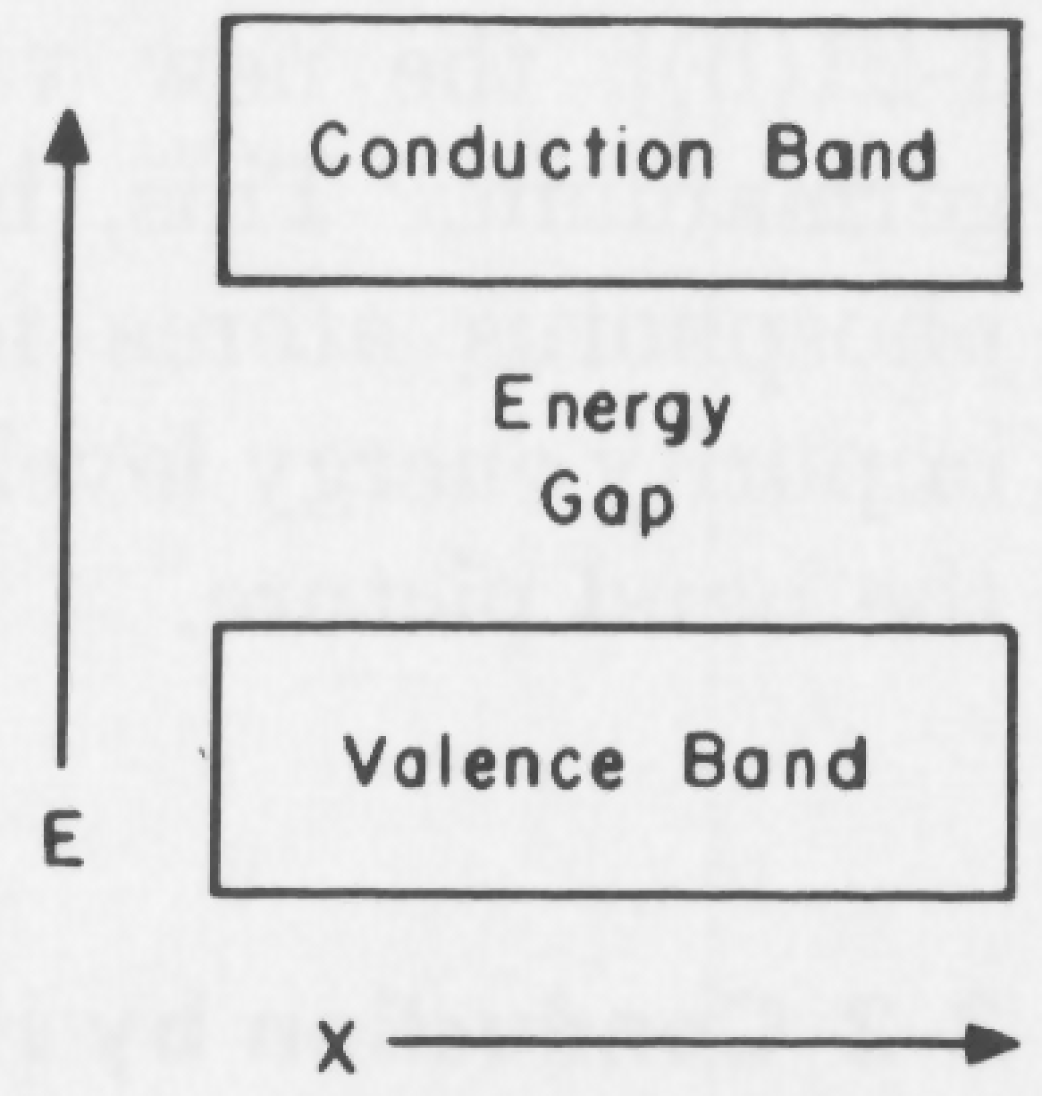


FIG. 3-4. The dependence of band structure on distance in a semiconductor.

The next thing to consider is the energy associated with a donor atom and the way in which this energy can be represented on an energy band diagram like Fig. 3-4. For a phosphorus donor atom, four electrons are involved in the covalent bonds, and the fifth one is free. The excess electron should act approximately as does the single electron of the hydrogen atom with the modification that it is in a germanium crystal whose dielectric constant $\epsilon_r = 16$. By Eq. (1-98), the binding energy of the electron in hydrogen depends upon $1/\epsilon^2$, and in germanium is

$$-13.5/16^2 = -0.05 \text{ eV}$$

Equation (1-98) also shows that the binding energy depends upon m. If we replace m by the approximate value of the effective mass, $m_e = 0.2\,m$, for electrons in germanium, we get an impurity ionization energy of -0.01 eV, which agrees with experiment. The value of 0.01 eV represents the binding energy of the excess electron, and this is, therefore, the energy needed to place it in the conduction band. This means that the energy level for this electron normally lies 0.01 eV below the bottom of the conduction band, that is, in the energy gap. In Fig. 3-5, we show the position of this level, with typical energy values for germanium. If we initially take the bottom of the conduction band as the level of zero energy, then the two values indicated in Fig. 3-5 are both negative. The impurity energy level is shown as a dashed line to indicate that the donor electrons are located reasonably close to the impurity atoms. While the radius of the hydrogen atom, by Eq. (1-98), is only 0.5 A [see Problem 1-21(b)], the new values of ϵ_r and m_e make it approximately 40 A in germanium. This, however, is small compared to the spacing between phosphorus atoms for even fairly impure germanium. Note that the impurity energy levels are distinct and we show one level per donor atom in the band picture.

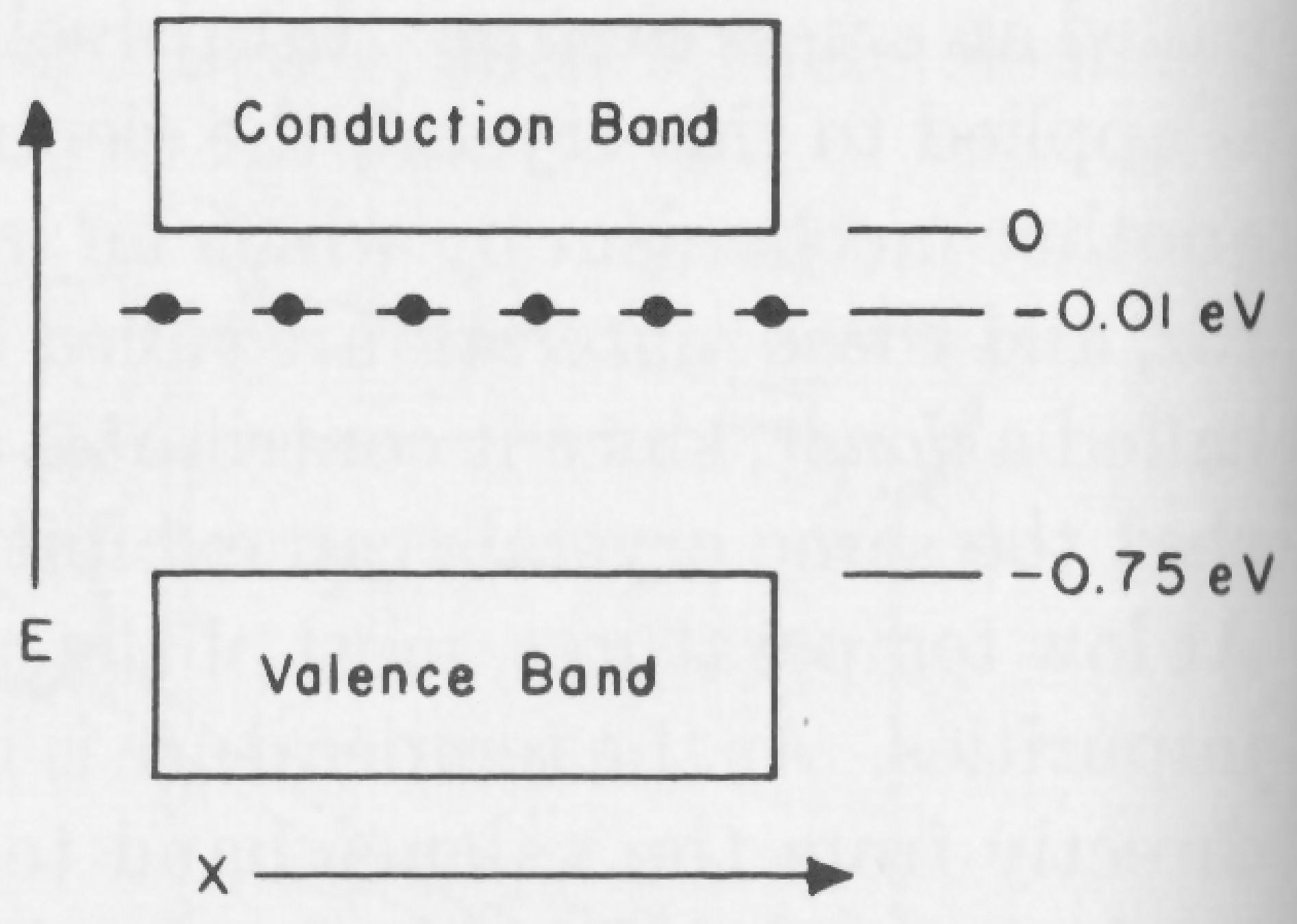


FIG. 3-5. The location of a donor level in the forbidden gap of germanium.

3-2 Conduction by Holes

Let us consider now what happens if we replace a germanium atom with a boron atom in a crystal. By Table 3-1, boron has only three valence electrons, so one of the covalent bonds will have an electron missing. If a potential is applied across the crystal, electrons are urged towards the positive terminal and one near the boron atom is able to drop into the vacancy in the covalent bond. This creates a new vacancy, and this process repeats itself until the vacancy eventually arrives at the negative terminal of the crystal. The vacancy created by a missing electron thus acts like a positive charge, since it moves in the direction opposite to that of a free electron from a donor atom in germanium. For this reason, this incomplete covalent bond is called a *positive hole*, or simply a *hole*. We can form an intuitive picture of the concept of a hole by considering the following analogy (which, like all analogies, should not be pressed too far):

Two warehouses, each 10 meters on a side, are used to store identical boxes of goods, which measure 1 by 1 meter at the base. One warehouse contains ninety-nine boxes and the other contains just one box, as shown in Fig. 3-6a.

PROBLEM 3-4

(a) One group of men shifts the boxes in warehouse number 1 so that the empty space moves from corner A to corner B. A similar group of men in warehouse number 2 moves the lone box from corner A to corner B. Which group does more work?

(b) Now convert the boxes into electrons, as shown in Fig. 3-6b, and apply potentials with polarities as shown to the walls of the two warehouses. In what direction does the box in warehouse number 2 move? Answer the same question for the empty space in warehouse number 1. What is the effective charge of this empty space?

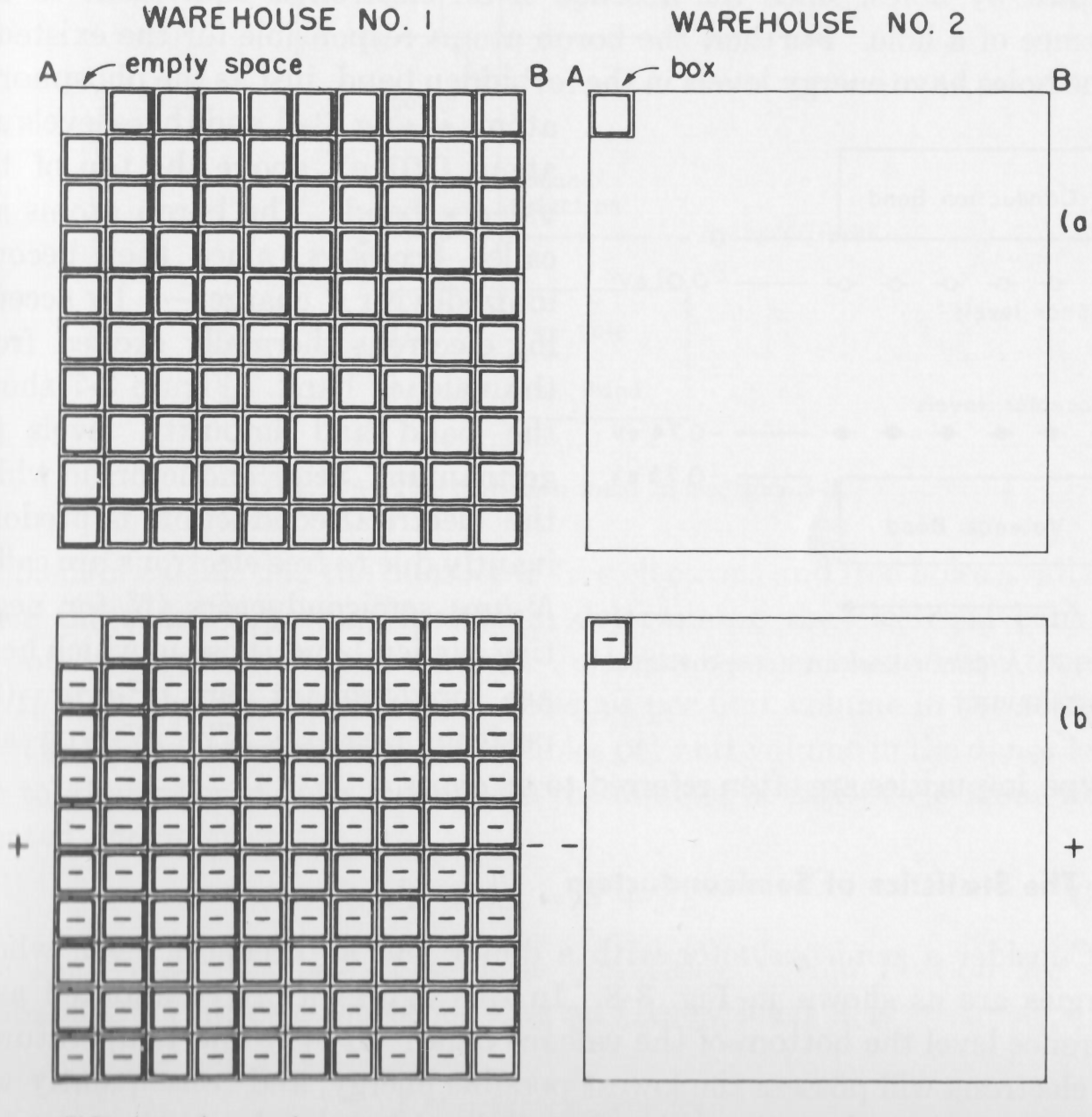


FIG. 3-6. A warehouse used as an analogy to explain hole conduction.

The empty space of this problem is like a hole, and it illustrates that a hole should have the same mass as an electron but a positive charge of $+1.6 \times 10^{-19}$ coulomb. When considering the motion of electrons in solids, we should deal with effective, rather than free, masses. Returning to Fig. 1-22, we see that at the *top* of an allowed band, d^2E/dk^2 is negative. This would appear to indicate that the effective mass m_h of the hole is negative, by Eq. (1-116). Therefore, we define m_h by:

$$m_h = \frac{-\hbar^2}{(d^2E/dk^2)} \tag{3-6}$$

By the energy band picture of Fig. 3-2, removal of an electron from the valence band creates an unoccupied level (if we neglect spin). Hence, it is possible for the electrons in the valence band to accept energy from an applied electric field, and a semiconductor whose valence band contains empty levels is changed from an insulator to a conductor. These empty levels, which correspond to missing electrons, therefore represent levels *occupied* by holes, since the absence of an electron is equivalent to the presence of a hole. Further, the boron atoms responsible for the existence of the holes have energy levels in the forbidden band, just as the phosphorus atoms in Fig. 3-5, and these levels are about 0.01 eV above the top of the valence band. The boron atoms are called *acceptors*, since they become ionized with a charge $-e$ by accepting electrons thermally excited from the valence band. Figure 3-7 shows the band and impurity levels for germanium. Semiconductors in which the electrical conduction is predominantly due to free electrons are called *N-type* semiconductors (N for negative carriers) and those in which holes are predominant are *P-type* (for positive carriers). The N-type and P-type impurities are often referred to as *dopes.*

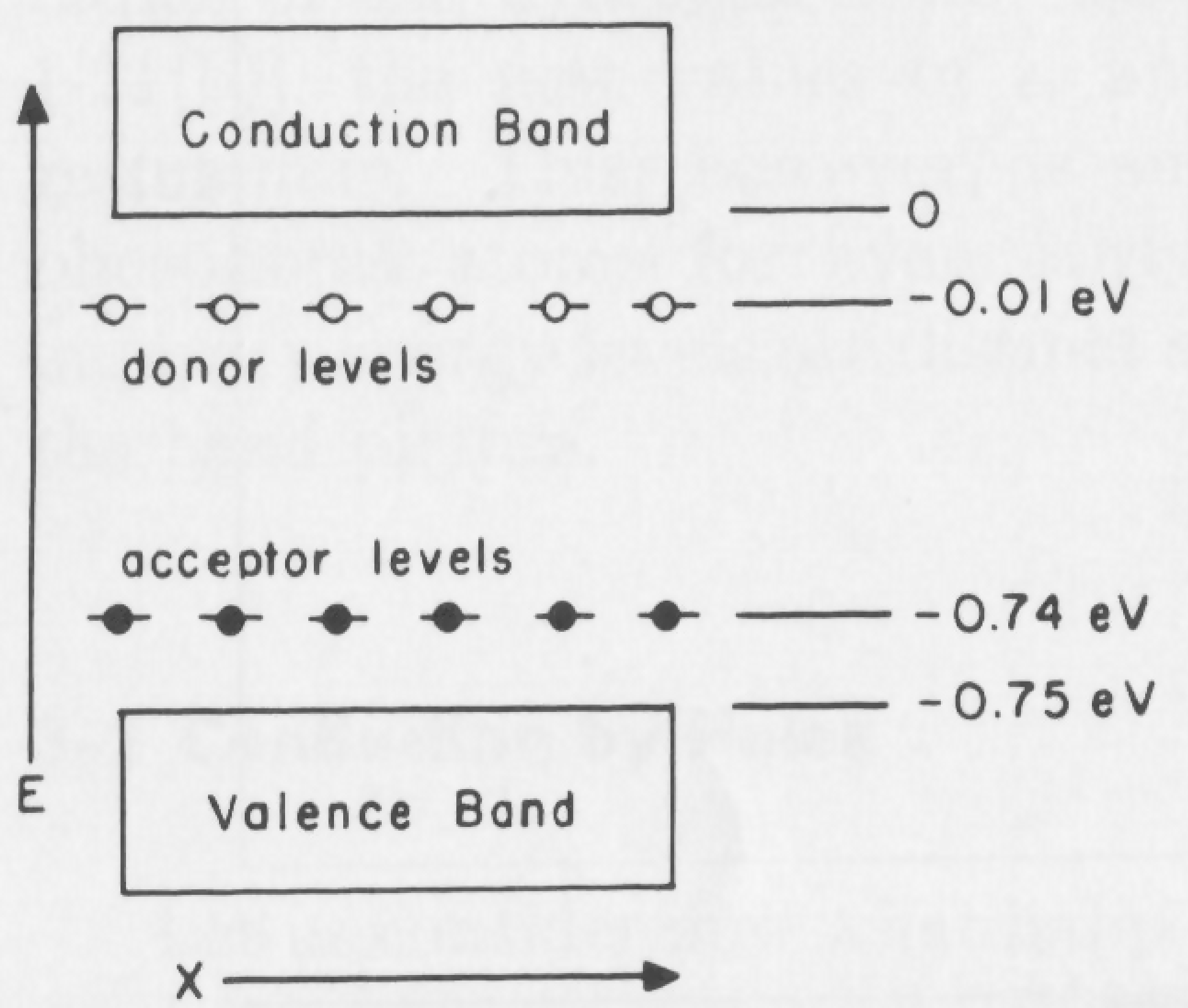


FIG. 3-7. A donor and an acceptor level in germanium.

3-3 The Statistics of Semiconductors

Consider a semiconductor with a donor and an acceptor level, whose energies are as shown in Fig. 3-8. In this figure, we have adopted as a reference level the bottom of the valence band. At very low temperatures, the electrons will possess the lowest possible energy, and consequently will all be located in the valence band and the donor level. As we raise the

temperature, we first excite electrons into the acceptor level from the valence band and into the conduction band from the donor level. At higher temperatures, electrons are able to get to the conduction band directly from the valence band. Electrons leaving the valence band for the acceptor level leave behind *free holes* and introduce *bound electrons* into the acceptor level. These processes are shown by arrows in Fig. 3-8. Similarly, the electrons leaving the donor level leave behind bound electrons. If p is the number of holes per unit volume, or the *carrier density,* in the valence band, we do not know what fraction of these p holes were created by electrons going to the acceptor level and what fraction by electrons going all the way to the conduction band. A similar statement holds for the n electrons per unit volume in the conduction band. We now set ourselves the problem of calculating the number of free electrons and free holes available, since these are important factors in determining the electrical behavior of both the semiconductors and the junction devices made from them.

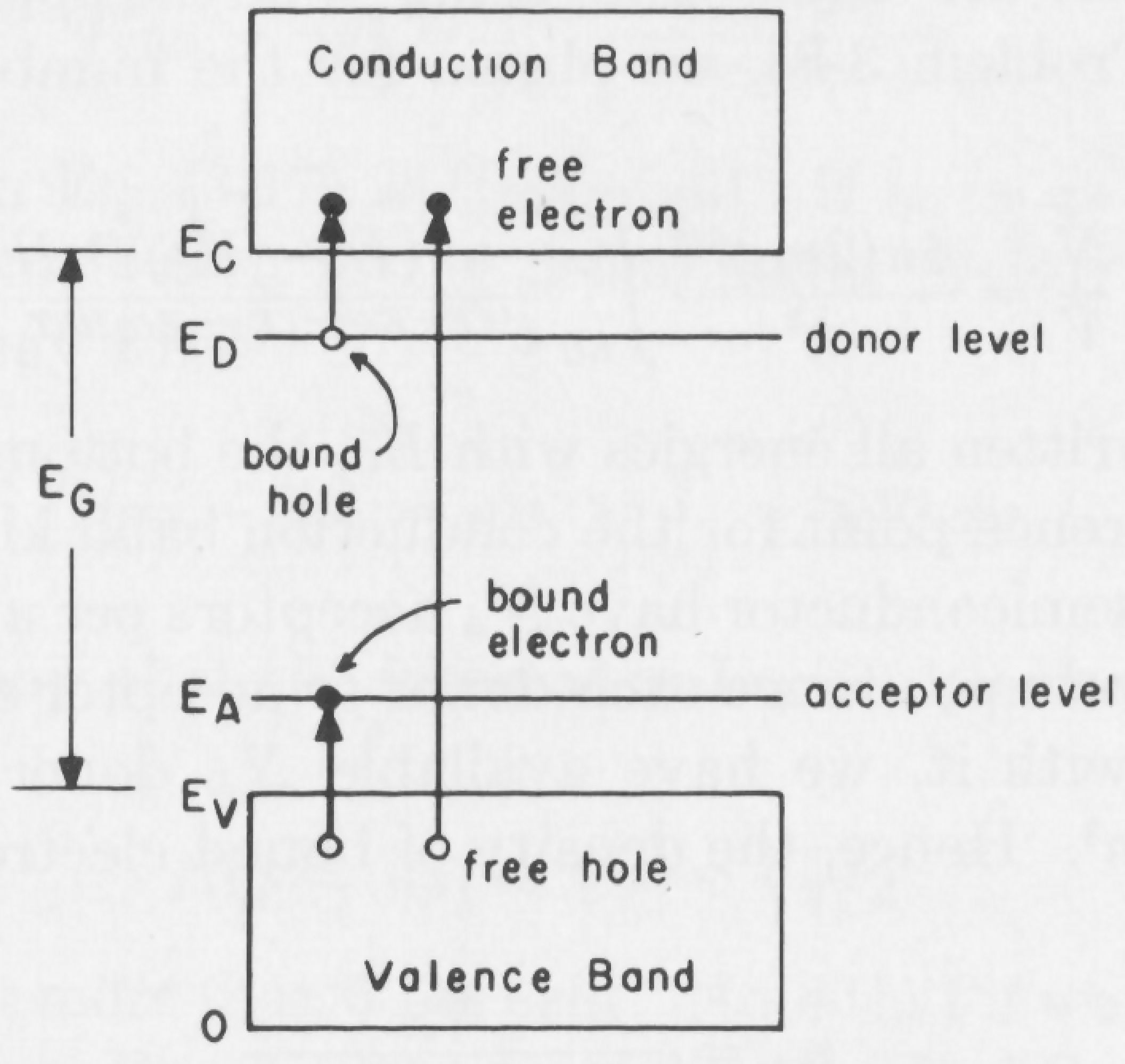


FIG. 3-8. The notation used in Section 3-3.

If n_A is the number of bound electrons per unit volume in the acceptor level and p_D is the number of bound holes per unit volume in the donor level, the total number of holes must equal the number of excited electrons which created them, or

$$n + n_A = p + p_D \tag{3-7}$$

We shall now calculate these four quantities. From Section 2-8, the probability of an electron occupying an energy level E is

$$f = \frac{1}{[e^{(E-E_F)/kT} + 1]}$$

The density of levels available in the lower part of the conduction band, where the electrons have an effective mass m_e, comes from Eq. (1-88).

$$dG = 4\pi(2m_e)^{3/2}E^{1/2}V\,dE/h^3 \tag{3-8}$$

where we have included a factor of 2 for spin. Since the lowest energy in the conduction band is E_C, we must modify Eq. (3-8) by calculating the points in the positive octant whose center lies at the point E_C. Hence,

$$dG = 4\pi(2m_e)^{3/2}(E - E_C)^{1/2}V\,dE/h^3 \tag{3-9}$$

Multiplying the probability of occupying a level by the density dG gives the number of electrons in the lower part of the conduction band as

$$dN = f\,dG$$

If we integrate over the entire band, using ∞ as the upper limit (which will be justified by Problem 3-8), we obtain for the number of electrons per unit volume, n,

$$n = \frac{N}{V} = \frac{4\pi(2m_e)^{3/2}}{h^3}\int_{E_C}^{\infty}\frac{(E - E_C)^{1/2}\,dE}{e^{[(E-E_C)-(E_F-E_C)]/kT} + 1} \tag{3-10}$$

where we have written all energies with E_C, the bottom of the conduction band, as the reference point for the conduction band kinetic energies.

Now, let the semiconductor have N_A acceptors per unit volume and N_D donors per unit volume. Since each donor or acceptor atom has a distinct level associated with it, we have available N_D donor levels/m^3 and N_A acceptor levels/m^3. Hence, the density of bound electrons in the acceptor level is

$$n_A = \frac{N_A}{e^{(E_A-E_F)/kT} + 1} \tag{3-11}$$

To compute hole densities, we realize that if f is the probability of an energy level being occupied by an electron, $(1 - f)$ is the probability that it will not be occupied, which is the same as saying that $(1 - f)$ is the probability of a level being occupied by a hole, or

$$1 - f = \frac{1}{e^{(E_F-E)/kT} + 1} \tag{3-12}$$

Hence, the density of holes in the donor level is

$$p_D = \frac{N_D}{e^{(E_F-E_D)/kT} + 1} \tag{3-13}$$

Finally, the density p of holes in the conduction band, if we use arguments similar to those for Eq. (3-10), is

$$p = \frac{4\pi(2m_h)^{3/2}}{h^3}\int_{-\infty}^{E_V}\frac{(E - E_V)^{1/2}\,dE}{e^{(E_F-E)/kT} + 1} \tag{3-14}$$

The integrals of Eqs. (3-10) and (3-14) can be put into the same form by letting

$$\left.\begin{aligned} \eta &= (E - E_C)/kT \qquad \lambda = (-E + E_V)/kT \\ \eta_C &= E_C/kT, \qquad \eta_V = E_V/kT, \qquad \eta_F = E_F/kT \end{aligned}\right\} \tag{3-15}$$

Then

$$n = \frac{4\pi(2m_ekT)^{3/2}}{h^3}\int_0^{\infty}\frac{\eta^{1/2}\,d\eta}{e^{(\eta+\eta_C-\eta_F)} + 1} \tag{3-16}$$

$$p = \frac{4\pi(2m_hkT)^{3/2}}{h^3}\int_0^{\infty}\frac{\lambda^{1/2}\,d\lambda}{e^{(\eta_F+\lambda-\eta_V)} + 1} \tag{3-17}$$

If we denote the integral in Eq. (3-16) by

$$I(\eta_C - \eta_F) = \int_0^{\infty}\frac{\eta^{1/2}\,d\eta}{e^{(\eta+\eta_C-\eta_F)} + 1} \tag{3-18}$$

then the integral in Eq. (3-17) is $I(\eta_F - \eta_V)$. If $\eta_C - \eta_F > 2$, then for all values of η, the exponential in the denominator of Eq. (3-18) is a large number, and we may write

$$I(\eta_C - \eta_F) = e^{\eta_F-\eta_C}\int_0^{\infty} e^{-\eta}\eta^{1/2}\,d\eta \tag{3-19}$$

Standard tables show that the integral in Eq. (3-19) has the value $\sqrt{\pi}/2$. Hence,

$$I(\eta_C - \eta_F) = e^{\eta_F-\eta_C}\sqrt{\pi}/2$$

with an error of not more than 5 per cent. Note that if we can neglect the 1 in the denominator of Eq. (3-18), we are changing δ of Eq. (2-59) from 1 to 0. In other words, this approximation is equivalent to using the Maxwell-Boltzmann distribution. In semiconductors at reasonable temperatures and with reasonable amounts of impurities, the Boltzmann distribution is generally quite a good approximation. For very high impurity concentration, the semiconductor is said to *degenerate* into a metal, and we must use the Fermi-Dirac law. Note that this approximation requires that the Fermi level E_F lie at least $2kT$ eV below the bottom of the conduction band. Similarly, if we require E_F to lie $2kT$ eV above the top of the valence band, an equivalent expression is obtained for the integral in Eq. (3-17). Using the abbreviations

$$N_C = 2(2\pi m_ekT)^{3/2}/h^3 \tag{3-20}$$

$$N_V = 2(2\pi m_hkT)^{3/2}/h^3 \tag{3-21}$$

we get

$$n = N_C\,e^{(E_F-E_C)/kT} \tag{3-22}$$

$$p = N_V\,e^{(E_V-E_F)/kT} \tag{3-23}$$

The product of n and p is denoted by n_i^2

$$np = n_i^2 = N_C N_V \, e^{(E_V - E_C)/kT} \tag{3-24}$$

The quantity n_i is called the *intrinsic carrier density.*

PROBLEM 3-5 Justify the name applied to n_i by the use of Fig. 3-8 and Eq. (3-7) for a perfectly pure semiconductor.

The quantity N_C in Eq. (3-18) is called the *effective density of states in the conduction band.* Since $e^{E_F - E_C/kT}$ has the form of a Maxwell-Boltzmann probability, it seems plausible that the coefficient by which we multiply it to obtain the number of electrons should be the density of states, in analogy with $dG = f\,dN$. Similarly, N_V is the effective density of states in the valence band.

The quantity $E_V - E_C$ in Eq. (3-24) is $-E_G$, as shown in Fig. 3-8. Hence,

$$n_i = \sqrt{N_C N_V} \, e^{-E_G/2kT} \tag{3-25}$$

Let us now define an energy level E_I, called the *intrinsic level,* by

$$E_I = \frac{E_C + E_V}{2} + \frac{kT}{2} \ln \left(\frac{N_V}{N_C} \right) \tag{3-26}$$

Note that when $m_e = m_h$, Eq. (3-26) reduces to

$$E_I = \frac{E_C + E_V}{2} = \frac{E_G}{2} + E_V \tag{3-27}$$

Hence, E_I lies approximately at the center of the forbidden gap, as shown in Fig. 3-9. Adding $-E_C$ to both sides of Eq. (3-26) gives

$$E_I - E_C = \frac{-E_C + E_V}{2} + \frac{kT}{2} \ln \left(\frac{N_V}{N_C} \right)$$

or

$$-E_C = -E_I + \frac{(E_V - E_C)}{2} + \frac{kT}{2} \ln \left(\frac{N_V}{N_C} \right) \tag{3-28}$$

By Eq. (3-22),

$$n = N_C \sqrt{N_V/N_C} \, e^{[E_F - E_I + (E_V - E_C)/2]/kT}$$

and by Eq. (3-25) this becomes

$$n = n_i \, e^{(E_F - E_I)/kT} \tag{3-29}$$

Similarly, it can be shown that

$$p = n_i \, e^{(E_I - E_F)/kT} \tag{3-30}$$

In Section 2-8, it was shown that

$$\bar{\mu} = E_F \tag{2-98}$$

and we defined the electrochemical potential $\bar{\mu}$ as

$$\bar{\mu} = K - e\mathcal{V} \tag{2-95}$$

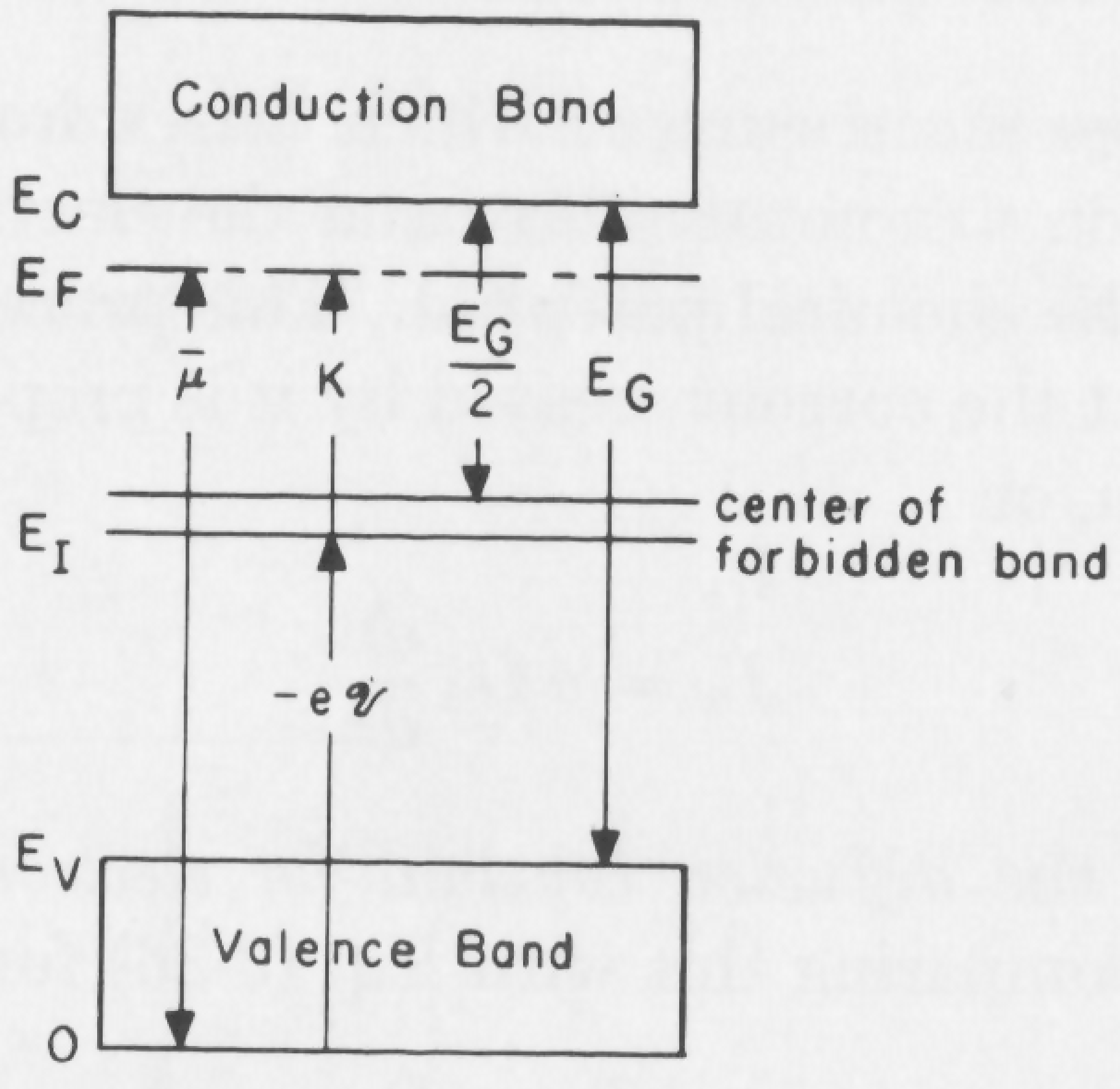


FIG. 3-9. The connection between the quantities of Figure 3-8 and of Chapter II.

If we now identify E_I with $-e\mathcal{V}$, then

$$K = \bar{\mu} + e\mathcal{V} = E_F - E_I \tag{3-31}$$

and Eqs. (3-29) and (3-30) become

$$n = n_i e^{K/kT} \tag{3-32}$$

$$p = n_i e^{-K/kT} \tag{3-33}$$

so that

$$np = n_i^2 \tag{3-34}$$

agreeing with Eq. (3-24). Using Eq. (3-34), we see that Eqs. (3-32) and (3-33) can be combined to give

$$K = kT \ln (n/n_i) = kT \ln (n_i/p) \tag{3-35}$$

It should be emphasized that this holds only for equilibrium, i.e., uniform temperature, pressure, and electrochemical potential. Note in Fig. 3-9 that K represents the distance from E_F to E_I ,in accordance with Eq. (3-31).

Equations (3-32) and (3-33) enables us to get some further insight into the meaning of the electrochemical potential $\bar{\mu}$. Let us consider an N-type semiconductor and calculate the electron current density J_n. By Eq. (3-32)

$$\frac{dK}{dx} = (kT/n) \frac{dn}{dx}$$

and from Eq. (1-24)

$$\sigma/e = n\mu_n$$

where μ_n is the electron mobility, so Eq. (2-102) becomes

$$J_n = \mu_n kT \frac{dn}{dx} + n\, e\mu_n \mathcal{E} \tag{3-36}$$

The first term represents a current which arises from the fact that the electron concentration n is non-uniform and the electrons shift themselves around to equalize the chemical potential. This process is called *diffusion*; we would expect that the current created by it is proportional to the variation in concentration, or

$$J_n = e\, D_n \frac{dn}{dx} \tag{3-37}$$

where D_n is called the *diffusion constant for electrons* and Eq. (3-37) is called *Fick's law.* Comparing this with Eq. (3-36) for $\mathcal{E} = 0$ yields

$$\mu_n kT = eD_n \tag{3-38}$$

This is known as the *Einstein relation.* Thus

$$J_n = e\, D_n \frac{dn}{dx} + n\, e\mu_n \mathcal{E} \tag{3-39}$$

For holes, we can show in a similar fashion:

$$J_p = -e\, D_p \frac{dp}{dx} + p\, e\mu_p \mathcal{E} \tag{3-40}$$

It should be emphasized that the validity of Eq. (3-38) rests upon the fact that Eqs. (3-22) and (3-23) are acceptable approximations.

PROBLEM 3-6 Let $\mathcal{E} = 0$ in Eqs. (3-39) and (3-40) and explain the significance of the sign on the right-hand side of each equation.

We would like to conclude this treatment of semiconductor statistics by showing how the Fermi distribution function f is used in conjunction with the energy band diagram to obtain a physical picture of what happens in a semiconductor when we change either the temperature of the impurity concentration. In Fig. 3-10a we show the energy band diagram for a perfectly pure semiconductor at room temperature with the Fermi function superimposed upon it so that the energy scales of the two diagrams coincide.

PROBLEM 3-7

(a) Show that for an intrinsic semiconductor, $K = 0$, and

$$E_F = E_I \tag{3-41}$$

(b) What does $K = 0$ mean?

By Eq. (3-41), the Fermi level is at the approximate center of the gap, as we have shown in the figure, and the effective width of the region in which f goes from 1 to 0 is about $2\,kT$, or 0.05 eV, at $T = 300°$K, which is almost the same as the gap indicated. However, the "tails" of this transition region extend beyond these limits (which by Problem 2-15 are about $1/e$th

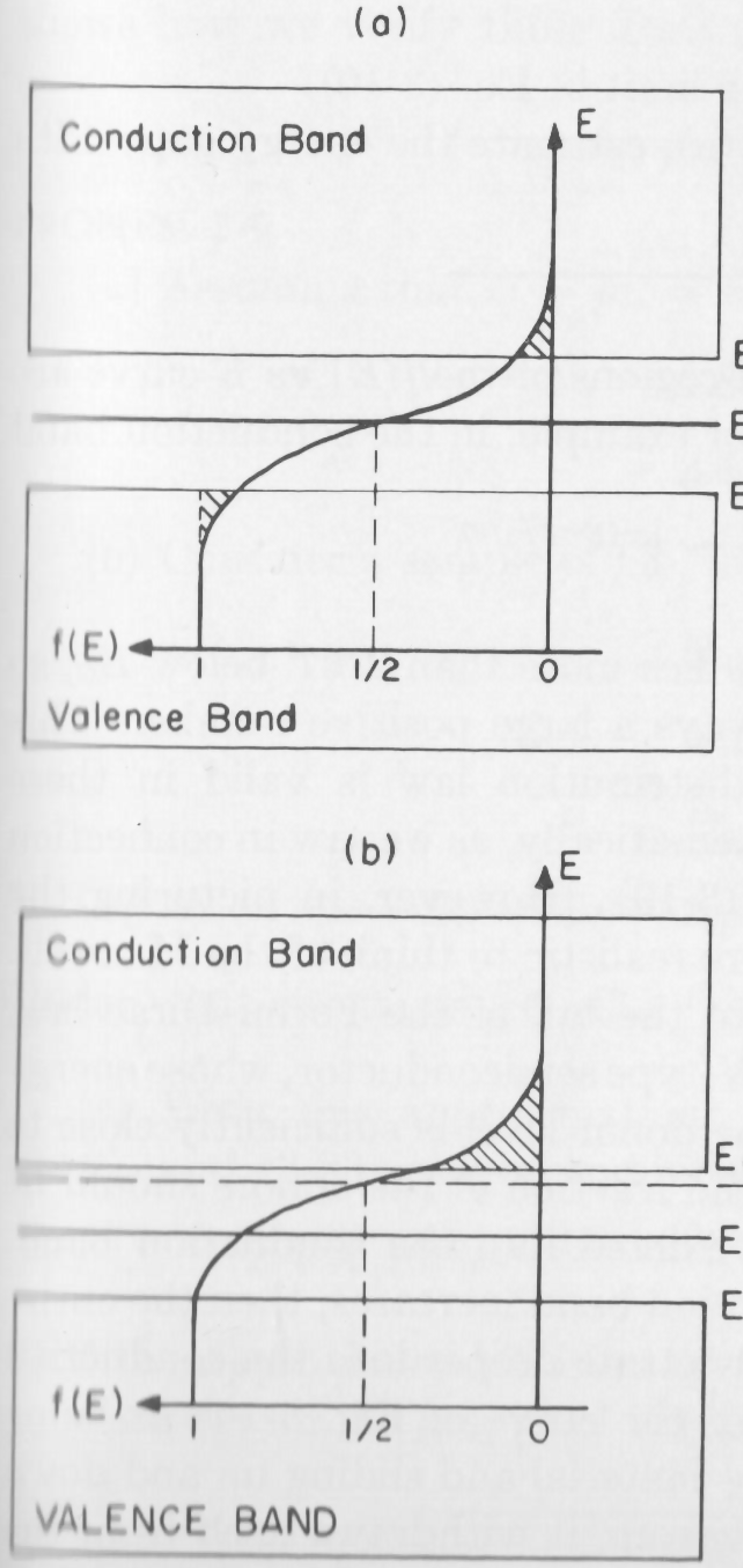


FIG. 3-10. The effect of the shift in the Fermi level on occupation of states in the valence and conduction bands. (The shape of the Fermi curve is exaggerated for clarity.)

of the way from $f = \frac{1}{2}$ to $f = 0$ or 1) and reach into the conduction and valence bands. In the valence band, for example, the probability of a level being occupied is 1, except very close to the top, where we have shaded the region where f begins to decrease. Then $(1 - f)$ in this region is small but not zero, and the levels cut by the shaded area are occupied by holes. Similarly, the shaded area in the conduction band corresponds to occupancy by electrons.

PROBLEM 3-8

(a) What happens to the shape of the $f(E)$ vs E curve when the temperature is raised? When it is lowered?

(b) How does this affect the shaded areas in Fig. 3-10 and hence the number of free electrons and holes?

(c) Justify the use of ∞ as the upper limit in Eq. (3-10).

(d) Using the result of Problem 2-16(a), estimate the energy gap width for the semiconductor in Fig. 3-10.

It is interesting to note that the tail regions of the $f(E)$ vs E curve are approximately exponential in shape. For example, in the conduction band

$$f = \frac{1}{e^{(E-E_F)/kT} + 1} \sim e^{-(E-E_F)/kT}$$

This approximation is valid because E_F lies more than 2 kT below E_C, so in the conduction band, $E - E_F$ is always a large positive number. This means that the Maxwell-Boltzmann distribution law is valid in these circumstances. This is convenient mathematically, as we saw in connection with the integrals of Eqs. (3-18) and (3-19). However, in picturing the behavior of electrons and holes, it is more realistic to think of the Maxwell-Boltzmann distribution as equivalent to the tail of the Fermi-Dirac law.

Consider next a moderately impure N-type semiconductor, whose energy band diagram is shown in Fig. 3-8. The donor level is sufficiently close to E_C, so at room temperature, a substantial fraction of the donors should be *ionized* — that is, have their electrons exicted into the conduction band. If the number of electrons in the conduction band increases, then the entire $f(E)$ vs E curve should move up and penetrate deeper into the conduction band (to visualize this process, think of the curve of Fig. 3-10a as being drawn on a piece of transparent overlay material and sliding up and down parallel to itself). When the curve moves up, it withdraws itself from the valence band, thus reducing the number of free holes. The Fermi level, of course, moves with it, so that E_F is closer to E_C in an N-type semiconductor than in an intrinsic one, as shown in Fig. 3-10b. The final position of

E_F depends upon the curve's crossing the three levels E_V, E_D, and E_C in such a way that the number of free holes in the valence band plus the number of bound holes in the donor level exactly equals the number of free electrons in the conduction band.

Suppose now the temperature is raised, changing the Fermi curve so that the transition region (the section where f goes from 1 to 0) is wider. Referring to Fig. 3-10b, we see that this has a much greater effect on the shaded area in the conduction band than in the valence band. The result would be to make the number of free electrons far exceed the number of free plus bound holes, and hence the curve and the Fermi level must slide down. On the other hand, if the number of donors is increased at a given temperature, the Fermi level would have to move farther up, just as it did when we added the original group of donor atoms. The problem below shows how we verify these ideas quantitatively.

PROBLEM 3-9

(a) Assuming that $m = m_e = m_h = 9.1 \times 10^{-31}$kg, show that

$$\begin{aligned} N_C = N_V &= 4.8 \times 10^{15}\, T^{3/2}\ \text{cm}^{-3} \\ &= 2.4 \times 10^{19}\ \text{cm}^{-3} \text{ at } 20°\text{C} \end{aligned} \tag{3-42}$$

(b) Consider a sample of Ge, for which

$$\begin{aligned} N_D &= 10^{15}\ \text{donors/cm}^3 \\ N_A &= 0 \\ E_C - E_D &= 0.01\ \text{eV} \\ E_G = E_C - E_V &= 0.75\ \text{eV} \end{aligned}$$

Under what circumstance will $e^{(E_F-E_D)/kT} \ll 1$?

(c) With this approximation, show that Eqs. (3-7), (3-11), (3-13), (3-22), and (3-23) can be combined to give

$$2.4 \times 10^{19}[e^{(E_F-E_C)/kT} - e^{(E_C-E_F-0.75)/kT}] - 10^{15} = 0 \tag{3-43}$$

Find the approximate position of E_F by solving for $e^{(E_F-E_C)kT}$ and check your answer on the graphs of Fig. 3-11.

By substituting a range of values of temperature for each value of N_D, we get the graphs of E_F vs T for germanium shown in Fig. 3-11. This figure brings out a feature of semiconductors which has not been explicitly mentioned before.

PROBLEM 3-10 Why do the bottom of the conduction band and the top of the valence band come closer together as the temperature gets higher in Fig. 3-11?

Note how the Fermi level actually gets up into the conduction band for very heavily doped (10^{19} donors/cm^3) materials. Similar calculations can be made for P-type germanium, and are also plotted in this figure. We shall show in a later chapter how the equations we have used here can be applied to the problem of measuring the individual donor and acceptor concentrations in semiconductors.

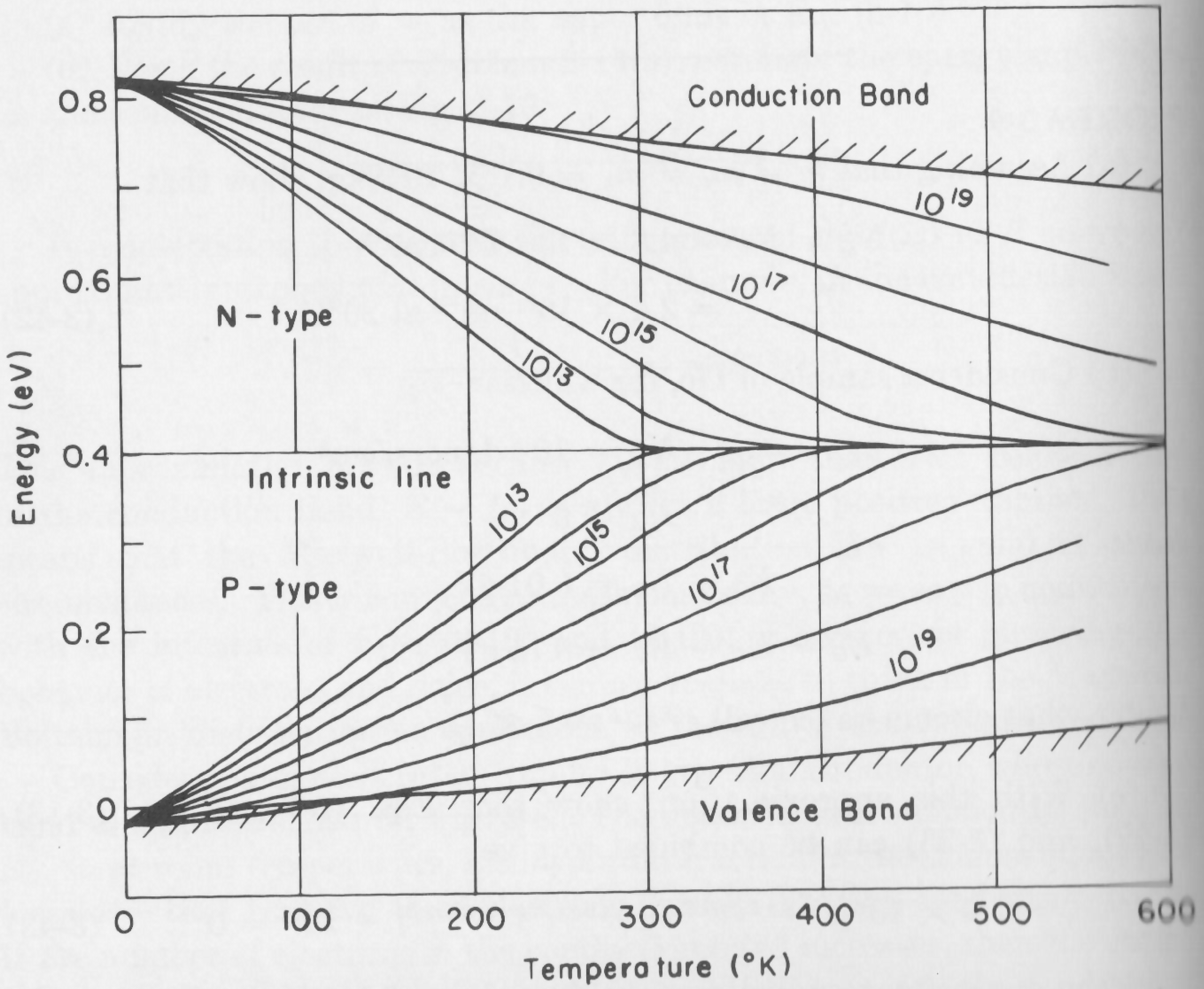


FIG. 3-11. Variation of the Fermi level with temperature for germanium. (From A. K. Jonscher, *Principles of Semiconductor Device Operation*, John Wiley, New York, 1960, with permission.)

3-4 The *PN*-Junction

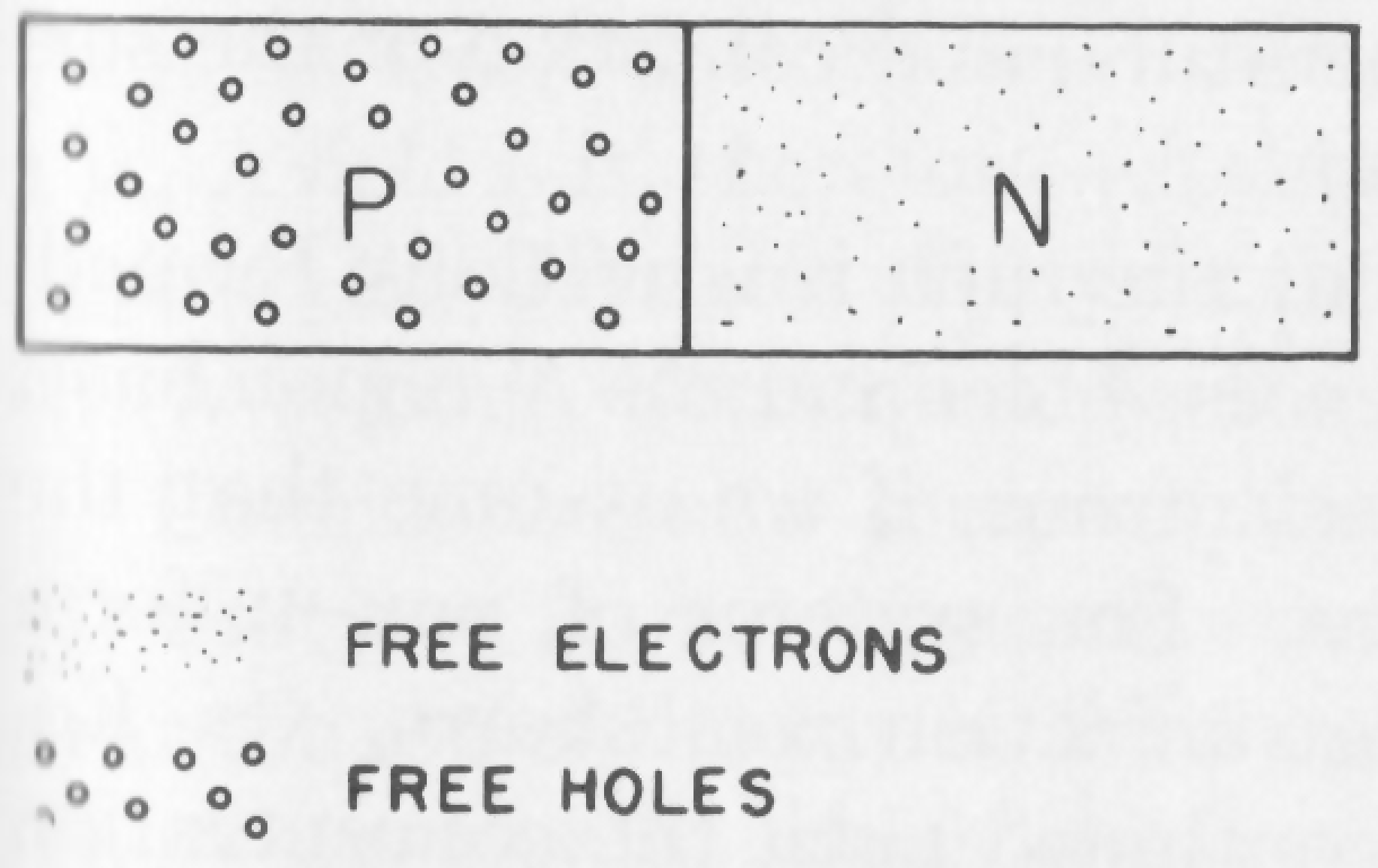


FIG. 3-12. A simple representation of a PN-junction.

Having considered the energy levels and the carrier densities in a uniform semiconductor crystal, we shall now extend these results to a PN-junction, which consists of an N-type region joined to a P-type region, as shown in Fig. 3-12. Before studying the behavior of a junction, we should describe how it is made, since it cannot be formed by simply pressing a block of P-type material against a block of N-type material. The surfaces are discontinuities in the regular crystal structure and would act as barriers to electron and hole flow. Junctions can be formed by starting with a block of, say, N-type material and converting part of it to P-type by one of the following methods:

Alloying A pellet of a P-type dope such as indium is placed on top of N-type germanium (Fig. 3-13a). We then heat the block and pellet to 500°C for a few minutes, causing the indium and some of the germanium in contact with it to melt and form a liquid In-Ge mixture. Upon cooling, the germanium recrystallizes with its original lattice arrangement, but since it is now heavily doped with indium it is P-type. The boundary between the original N-type crystal and the recrystallized (or *regrown*) P-type region is the PN-junction (Fig. 3-13d).

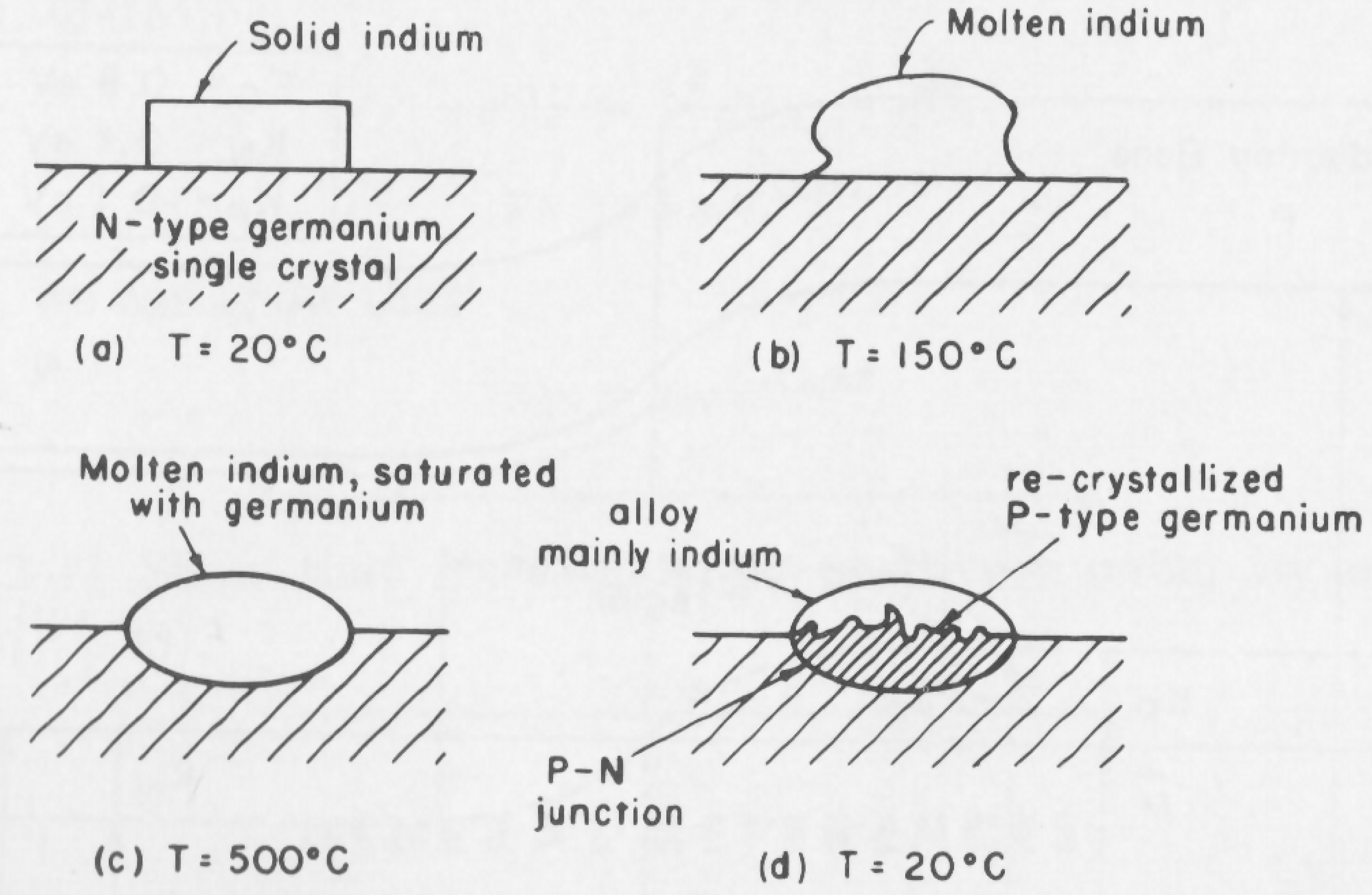


FIG. 3-13. The steps in the fabrication of an alloyed junction. (From E. W. Herold, Brit. Journ. Appl. Phys. 5, 115, 1954, with permission.)

Gaseous diffusion A gas such as BCl_3 is passed over a heated slab. The heated gas decomposes, and the resulting boron will diffuse into the crystal to a depth determined by the time, temperature, and condition of the surface, forming a *PN*-junction.

In Fig. 3-12, we show a *PN*-junction at the time it is initially formed. The *P* region has a high concentration of excess holes and the *N* region has a somewhat lower concentration of excess electrons, if we assume that the *N*-type starting material was fairly pure. For systems of non-uniform composition, the principle of entropy maximization was shown, via Eq. (2-18), to lead to a particle flow which continued until the concentration was everywhere uniform. When we are dealing with charged particles, it is not the chemical potential *K*, but the electrochemical potential $\bar{\mu}$ which must be everywhere uniform. Hence, holes will diffuse across the *PN*-junction into the *N*-type region, and electrons will go into the *P*-type region. The initial group of holes diffusing into the *N*-region from the neighborhood of one junction leave behind a layer of ionized acceptors, indicated in Fig. 3-14 as a row of negative charges. Similarly, the electrons leave behind a layer of positively charged acceptors. This *double layer* of charge is called the *space-charge* region or the *depletion layer* because it contains very few free electrons and holes. The width of this layer is usually about 1 μ. A field $\mathcal{E}$ is, therefore, established in the direction shown, and this field opposes the diffusion of any further electrons into the *P*-region or holes into the *N*-region. Consequently, the double layer is also called a *barrier layer*.

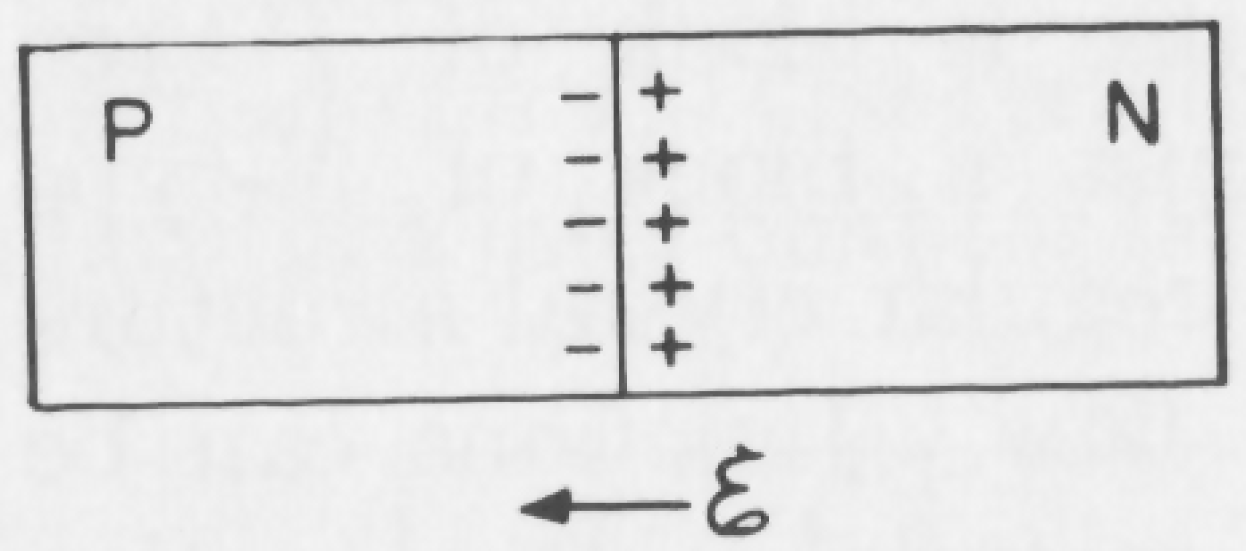


FIG. 3-14. The barrier layer in a *PN*-junction.

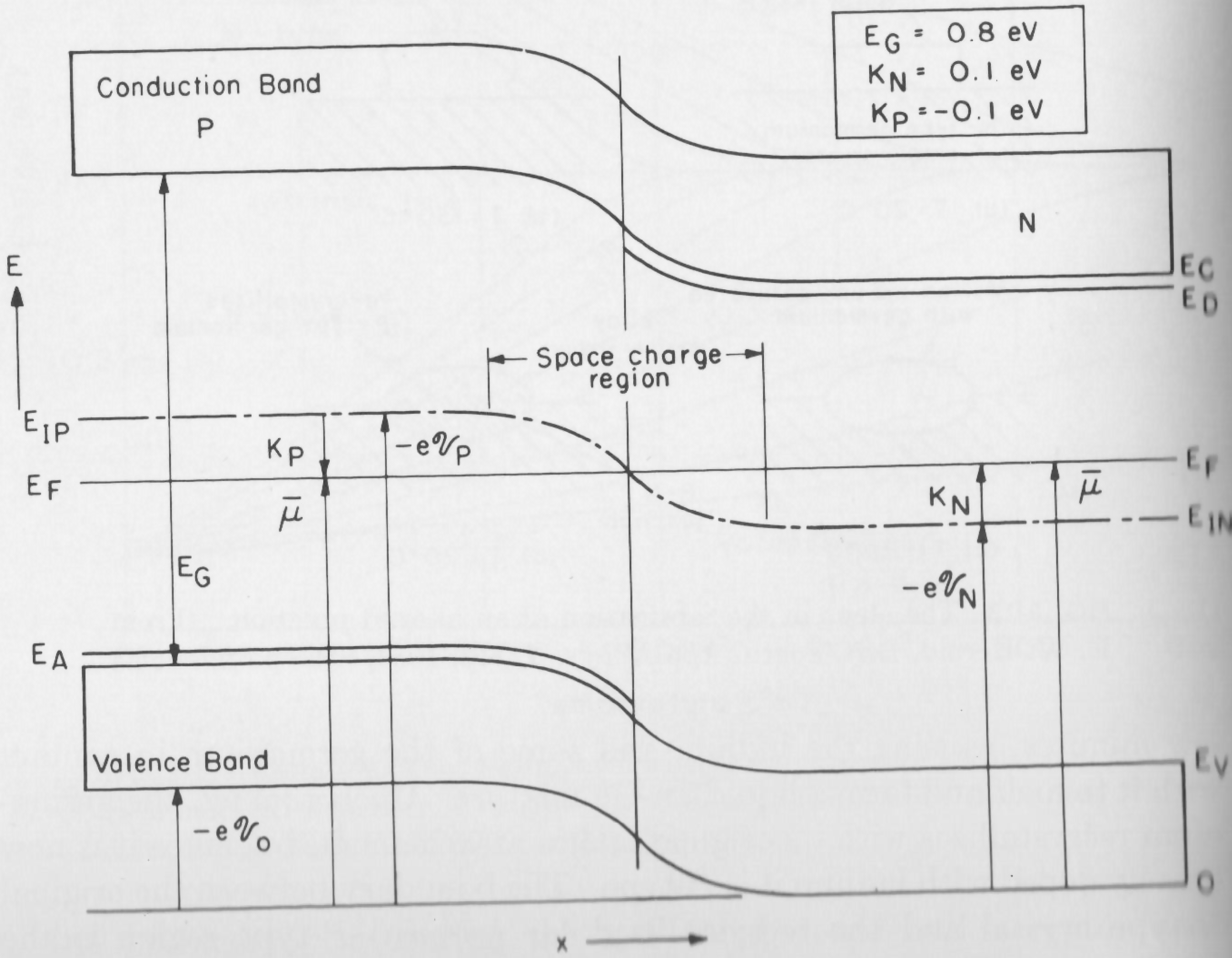


FIG. 3-15. The energy levels for the *PN*-junction in equilibrium.

Considering now the complete junction in equilibrium, we have shown (1) that $\bar{\mu}$ or E_F must be uniform over the entire junction and (2) that E_F is near the top of the gap in the *N*-region and near the bottom in the *P*-region. The only way in which these conditions can be simultaneously satisfied is for the bands to adjust themselves as shown in Fig. 3-15. The *N*-type region is taken as the reference, and the shift in the *P*-type bands is $-e\mathcal{V}_0$ where $\mathcal{V}_0$ is called the *barrier* or *built-in potential*. Note that $e\mathcal{V}_0$ represents the distance between the Fermi levels in the two sides of the junction before the *P*-band was raised. Since $\bar{\mu}$ is a constant, by Eq. (3-35),

$$kT \ln (n_i/p_N) - e\mathcal{V}_N = kT \ln (n_i/p_P) - e\mathcal{V}_P$$

where the subscripts refer to the *N* or *P* regions. Rewriting this gives

$$p_N = p_P\, e^{e(\mathcal{V}_P - \mathcal{V}_N)/kT}$$

From Fig. (3-15),

$$-e\mathcal{V}_P = -e\mathcal{V}_N - e\mathcal{V}_0$$

so

$$p_N = p_P\, e^{e\mathcal{V}_0/kT} \tag{3-44}$$

Similarly, we can show that

$$n_P = n_N\, e^{e\mathcal{V}_0/kT} \tag{3-45}$$

PROBLEM 3-11 Show that K_N is a large, positive number by using Eqs. (3-24) and (3-35).

GENERAL REFERENCES

Blakemore, J. S., *Semiconductor Statistics*. New York: Pergamon, 1962.

A more complete and precise account of many of the topics covered in this chapter will be found in this authoritative monograph. The discussions of some of the fine points of semiconductor physics are particularly good.

4

THE THEORY OF JUNCTION DEVICES

4-1 The PN-Junction in Equilibrium

At the end of Section 3-4, it was shown how the requirement that $\bar{\mu}$ be constant throughout a *PN*-junction in equilibrium enabled us to predict the way the energy bands vary with position. We should now like to calculate this variation in a fairly accurate way, and for this purpose, we will introduce *Poisson's equation.* We start by defining the *solid angle* ω of a cone as the surface area which the cone intercepts on a unit sphere. Consider a cone of infinitesimal base with area dA tilted at an angle θ, as shown in Fig. 4.1, where $d\mathbf{A}$ is a vector normal to the area dA and equal to it in magnitude. The projection of dA on a sphere of radius r is $dA \cos\theta$, and the solid angle which the cone subtends is given by

$$\frac{dA\cos\theta}{r^2} = \frac{d\omega}{1^2} \tag{4-1}$$

We define a *spherical radian* as the solid angle subtended by an area equal to r^2, so that there are $4\pi r^2/r^2$, or 4π, spherical radians to a sphere.

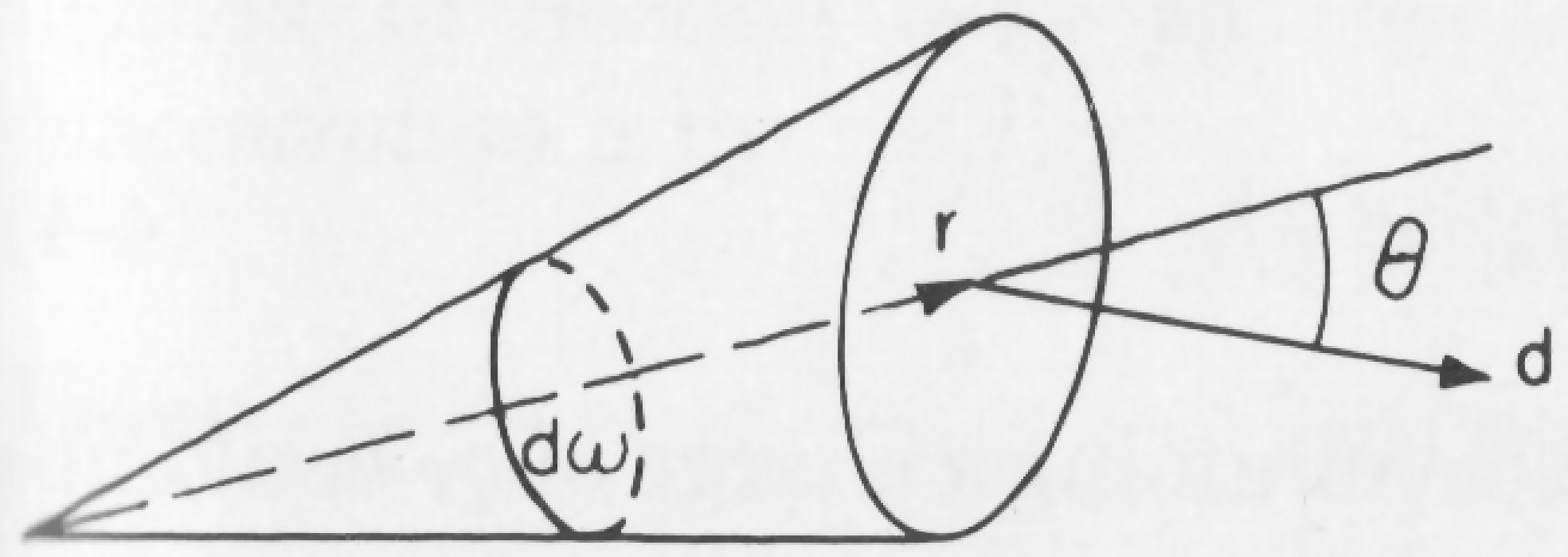


FIG. 4-1. The geometrical relation of an element of area dA to the solid angle $d\omega$ which it subtends.

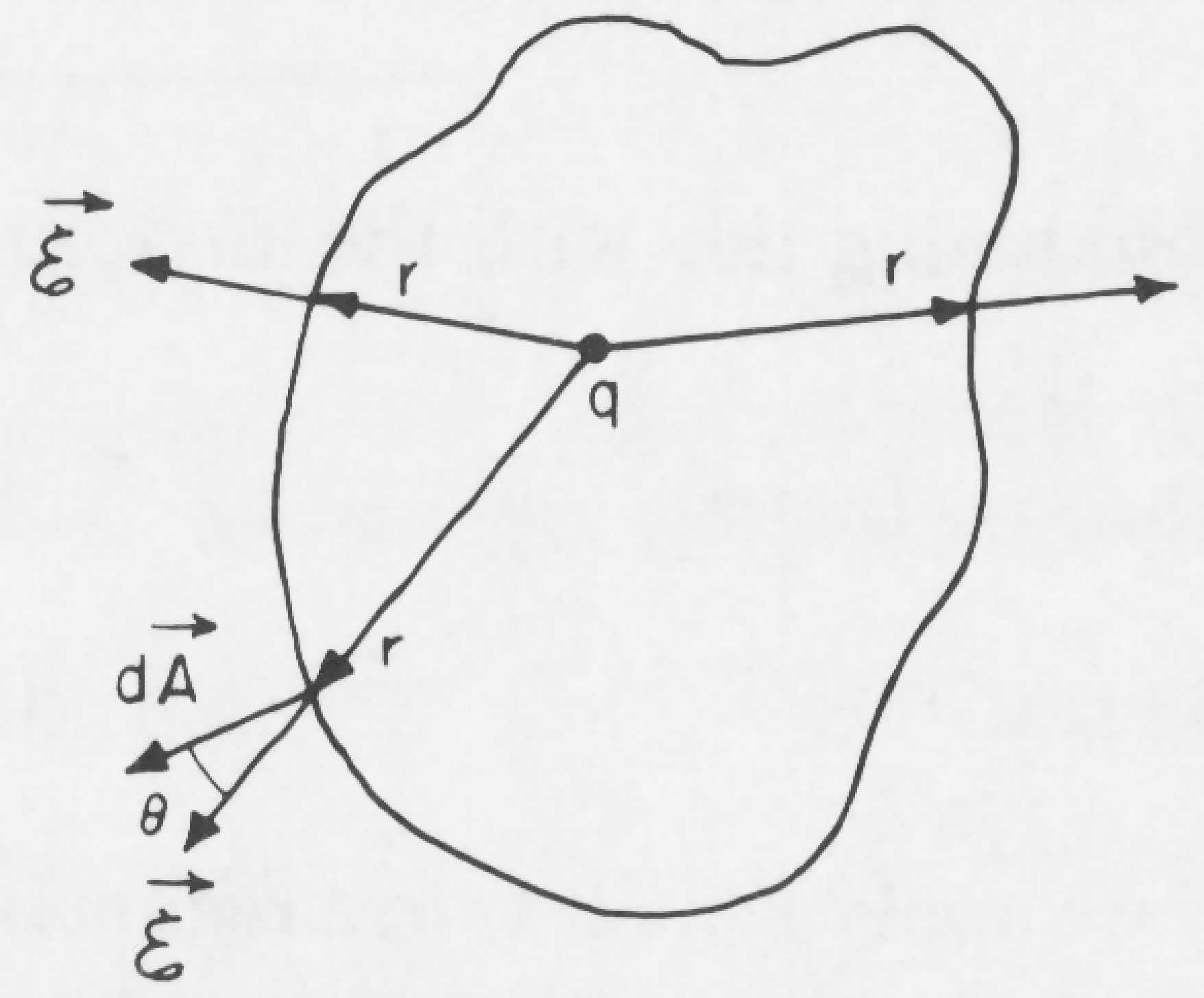


FIG. 4-2. The electrical field on a closed surface surrounding a charge q.

Now consider a charge q surrounded by an arbitrary closed surface and producing an electric field $\boldsymbol{\mathcal{E}}$ which varies from point to point over this surface, as shown in Fig. 4-2. The magnitude of $\mathcal{E}$ at each point on the surface, from Eqs. (1-12) and (1-14), is

$$\mathcal{E} = q/4\pi\epsilon r^2$$

and the component of $\boldsymbol{\mathcal{E}}$ normal to the surface is

$$\mathcal{E}_n = \mathcal{E}\cos\theta = \boldsymbol{\mathcal{E}}\cdot d\mathbf{A}/dA \tag{4-2}$$

since $d\mathbf{A}/dA$ is a vector of unit length. If we multiply $\mathcal{E}_n$ by dA and integrate over the entire surface, we obtain

$$\oint \mathcal{E}_n\, dA = \oint \left(\frac{q}{4\pi\epsilon r^2}\right)\cos\theta\, dA$$

where the symbol $\oint$ denotes integration over the entire surface; it is discussed in the appendix. By Eq. (4-2), this becomes

$$\oint \mathcal{E}_n\, dA = \frac{q}{4\pi\epsilon}\int_0^{4\pi} d\omega = \frac{q}{\epsilon} \tag{4-3}$$

This expression is called *Gauss' law.*

PROBLEM 4-1 Consider a uniformly charged conducting plane with a charge

q_A per unit area on its upper surface. Erect a right cylinder on this surface of area dA and show by Gauss' law that

$$\mathcal{E} = q_A/\epsilon$$

where $\mathcal{E}$ is the uniform field due to the charge.

Combining this with the *divergence theorem* (A46) given in the appendix:

$$\oint \boldsymbol{\mathcal{E}}\cdot d\mathbf{A} = \operatorname{div} \boldsymbol{\mathcal{E}}\, \mathrm{dV}$$

we get

$$\int \operatorname{div} \boldsymbol{\mathcal{E}}\, dV = q/\epsilon \tag{4-4}$$

If we apply Eq. (4-4) to a region in space containing a charge of q_V coulombs per unit volume, then it becomes

$$\int \operatorname{div} \boldsymbol{\mathcal{E}}\, dV = (1/\epsilon) \int q_V\, dV \tag{4-5}$$

Since this relation must hold for an infinitesimal volume dV, we equate the integrands to get

$$\operatorname{div} \boldsymbol{\mathcal{E}} = q_V/\epsilon \tag{4-6}$$

Equation (1-16), when combined with Eq. (1-5), is

$$\mathcal{V} = -\int \boldsymbol{\mathcal{E}}\cdot d\mathbf{r} \tag{1-16}$$

and the inverse of this, by (A41) of the appendix, is

$$\boldsymbol{\mathcal{E}} = -\mathbf{grad}\, \mathcal{V} \tag{4-7}$$

Using the definition (A31), we finally get

$$\nabla^2 \mathcal{V} = -q_V/\epsilon$$

which is Poisson's equation. We shall use this in the one-dimensional form

$$\frac{d^2\mathcal{V}}{dx^2} = \frac{-q_V}{\epsilon} \tag{4-8}$$

PROBLEM 4-2

(a) Using the idea that $\nabla^2\mathcal{V}$ is the "bulginess" in $\mathcal{V}$, as discussed in Appendix A3, give a physical interpretation of Poisson's equation as applied to a PN-junction.

(b) From the curve of $\mathcal{V}$ vs x shown in Fig. 3-15 and from Eq. (4-7), draw an approximate curve for $\boldsymbol{\mathcal{E}}$ vs x.

(c) From this new curve and Eq. (4-8), draw a curve of q_V vs x. Use this result to show that Poisson's equation leads to the space-charge layer of Fig. 3-14.

If all the impurities are ionized, the total charge density at each point due to both free and bound charges is

$$q_V = e(N_D - N_A + p - n) \tag{4-9}$$

where, as always, $e = +1.6 \times 10^{-19}$ coulomb; the other symbols are defined in Section 3-3. In this equation, define the *relative impurity concentration* a by

$$a = (N_D - N_A)/2n_i \tag{4-10}$$

Using

$$n = n_i\, e^{K/kT} \tag{3-32}$$

$$p = n_i\, e^{-K/kT} \tag{3-33}$$

and defining the *relative potential* u by

$$u = K/kT \tag{4-11}$$

gives

$$\frac{d^2\mathcal{V}}{dx^2} = \frac{2n_i e}{\epsilon} (\sinh u - a) \tag{4-12}$$

Using the definition

$$K = \bar{\mu} + e\mathcal{V} \tag{2-95}$$

and the fact that $\bar{\mu}$ is constant for equilibrium conditions, $\mathcal{V}$ can be eliminated from Eq. (4-12) to give

$$\frac{d^2u}{dx^2} = \frac{1}{L_D^2} (\sinh u - a) \tag{4-13}$$

where we have made the substitution

$$L_D = \sqrt{\epsilon\, kT/2n_i e^2} \tag{4-14}$$

The quantity L_D is called the *Debye length.* In silicon, for example, if we use $\epsilon_r = 12$ [Conwell (1952)[1]] and $n_i = 1.5 \times 10^{10}$ cm^3 at room temperature [Conwell (1958)[2]], the Debye length is about 0.001 cm.

PROBLEM 4-3 Calculate L_D for silicon.

We can see its physical significance by assuming that $K \ll kT$ and $(N_D - N_A) \ll n_i$.

PROBLEM 4-4

(a) What is the physical meaning of the two conditions above? Use Fig. 3-15 to assist in determining your answer.

(b) Show that an acceptable solution of Eq. (4-13) in the N-region under the assumptions mentioned above is

$$u = e^{-x/L_D}$$

and use it in conjunction with the figure to show that the effective width of the space-charge layer is about $2\,L_D$.

In solving Eq. (4-13) it is convenient to have the variables in dimensionless form, and we introduce a relative displacement r by

$$r = \sqrt{2a}\,x/L_D \tag{4-15}$$

following the treatment of Passau and Van Styvendael (1960)[3]. This converts Eq. (4-13) into

$$2a\,\frac{d^2u}{dr^2} = \sinh u - a \tag{4-16}$$

We see from Fig. 3-15 that the boundary conditions for a junction in equilibrium are

$$\frac{dK}{dx} = \frac{d^2K}{dx^2} = 0 \qquad \text{at } x = \pm\infty \tag{4-17}$$

Putting this into Eq. (4-13) yields

$$a = \sinh u_m \tag{4-18}$$

where u_m is the maximum value of u; in each of the two regions, u_m can be expressed

$$\left.\begin{aligned} u_m &= K_N/kT \qquad \text{for } x \gg 0 \\ u_m &= K_P/kT \qquad \text{for } x \ll 0 \end{aligned}\right\} \tag{4-19}$$

PROBLEM 4-5 Consider an N-type semiconductor with $N_A = 0$, with all the donors ionized, and with the Fermi level located so that Eqs. (3-22) and (3-23) are valid. Show that the position of the Fermi level is given by solving Eq. (3-7) agrees with its position as determined by Eq. (4-18).

Then Eq. (4-16) becomes

$$2\,\frac{d^2u}{dr^2} = \frac{\sinh u}{\sinh u_m} - 1 \tag{4-20}$$

If we assume that $N_D - N_A$ has a constant (although different) value in each half of the junction, Eq. (4-20) may be integrated once to give

$$\left(\frac{du}{dr}\right)^2 = \frac{\cosh u}{\sinh u_m} - u + C_1 \tag{4-21}$$

PROBLEM 4-6 Obtain Eq. (4-21) from Eq. (4-20) by letting $p = du/dr$ and integrating.

The boundary conditions [Eq. (4-17)] show that

$$C_1 = u_m - \coth u_m$$

Hence

$$\left(\frac{du}{dr}\right)^2 = \frac{\cosh u - \cosh u_m}{\sinh u_m} - (u - u_m) \tag{4-22}$$

or

$$r = \int_0^u \frac{du}{\pm\sqrt{\dfrac{\cosh u - \cosh u_m}{\sinh u_m} - (u - u_m)}} \tag{4-23}$$

The integral in this equation cannot be expressed in terms of known functions, but it is not difficult to evaluate numerically, as shown in the following problem.

PROBLEM 4-7

(a) Consider N-type germanium at room temperature with $N_D = 10^{15}$ donors/cm³ and $N_A = 0$. Using Fig. 3-11, show that Eq. (4-23) becomes

$$r = \int_0^u \frac{du}{\pm\sqrt{\dfrac{\cosh u}{23.5} - u + 2.85}} \tag{4-24}$$

(b) Show that Eqs. (4-10) and (4-18) agree numerically.

(c) Plot the integrand of Eq. (4-24) for $u = 0.5, 1.0, 1.5, \ldots$ using graph paper with quarter-inch squares. By counting squares, plot u as a function of r.

(d) Show that this curve of u vs r can be approximately fitted to an equation of the form

$$u = u_m\,(1 - e^{-\alpha r}) \tag{4-25}$$

where α is determined by the condition

$$u = u_m\,(1 - e^{-1}) = 0.63\,u_m$$

when

$$\alpha r = 1$$

(e) How is the curve of u vs r used in Fig. 3-15?

(f) Check the drawings that were made in Problem 4-2(b) and (c) for agreement with Eq. (4-25).

Our calculations so far have been restricted to the N-type region of a junction. To treat the complete junction, we must make a slight modification in the above procedure. It will be realized that the solution of the differential equation [Eq. (4-20)] involved a single arbitrary constant C_1, and that the other constant which we would expect from a second-order equation was eliminated by specifying the lower limit on the integral Eq. (4-23). This limit implies that $u = K = 0$ at $r = x = 0$, as shown in Fig. 3-15. However, it should be noted that we have chosen this junction to be *symmetric* and *abrupt;* symmetric because the Fermi level lies about the same distance below the intrinsic level at the end of the P-region as it does above the intrinsic level at the end of the N-region, and abrupt because the impurities change suddenly from P-type to N-type at the junction. Consider as an example a germanium PN-junction with impurity concentrations as follows:

$$N_D = 10^{15}\ \text{donors/cm}^3, \qquad N_A = 0 \qquad \text{for } x \geqq 0$$

$$N_A = 10^{18}\ \text{acceptors/cm}^3, \qquad N_D = 0 \qquad \text{for } x \leqq 0$$

From Fig. 3-11, we see that the maximum values of the chemical potentials are

$$K_{Nm} = 0.10\ \text{eV}, \qquad u_N = 0.10/0.026 = 3.85$$

$$K_{Pm} = -0.28\ \text{eV}, \qquad u_P = \qquad -10.8$$

If we sketch the band structure of this junction, we obtain Fig. 4-3, which shows that for this asymmetrical doping, the Fermi level lies below the center of the gap at the junction and the charges have redistributed themselves so that holes are the majority carrier at this point. In fact, the material is effectively P-type at $x = 0$, even though this is the location of the junction. This phenomenon produces an *inversion layer* in the N-type material, and the chemical potential K_{in} in this layer is negative, as it should be in a P-type region. Letting $K = K_0$ and $u = u_0$ at $r = 0$, we impose the boundary condition

$$\frac{d\,u_N}{dx} = \frac{d\,u_P}{dx} \tag{4-26}$$

at $u = u_0$.

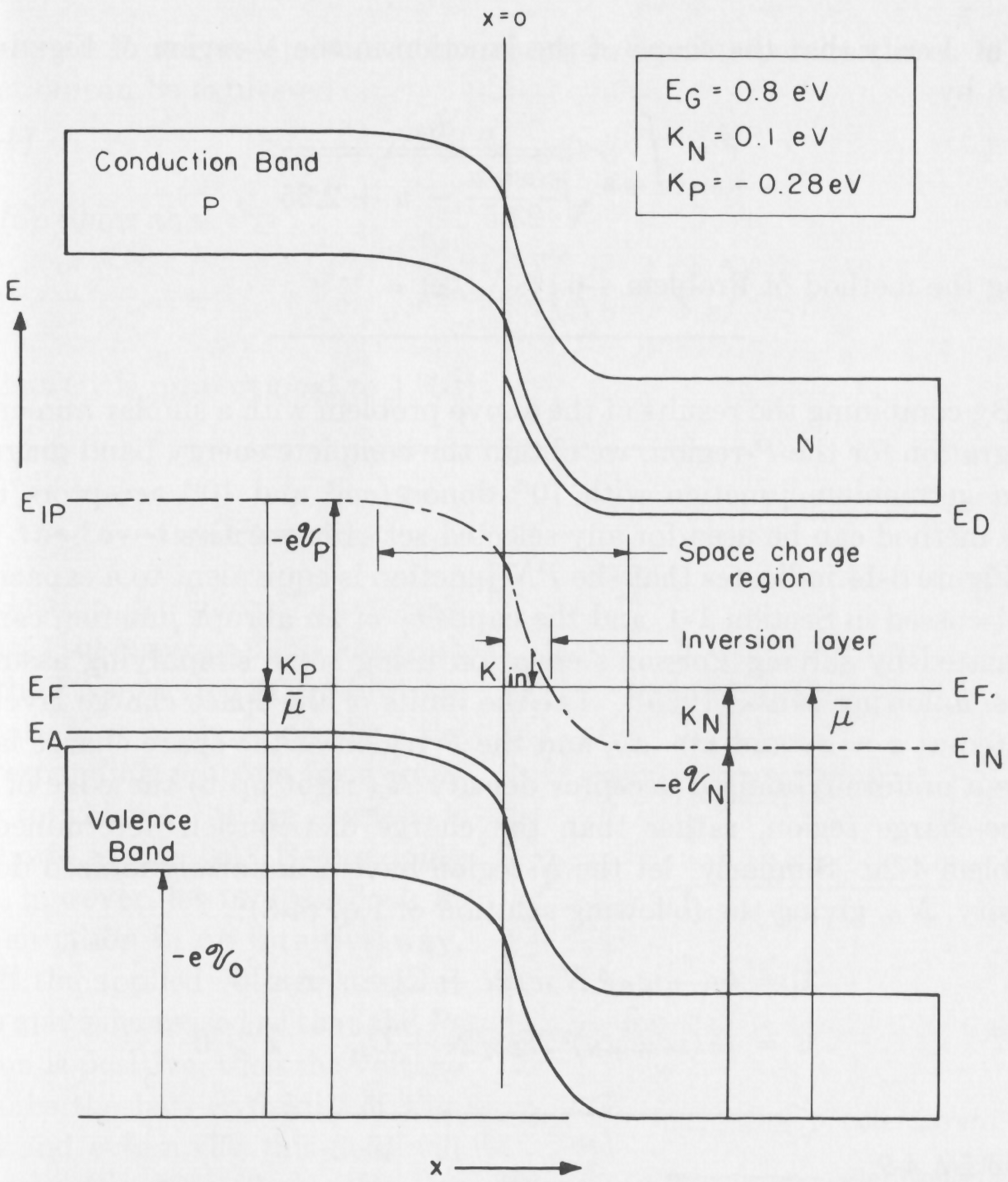


FIG. 4-3. The energy level diagram for the unsymmetrical PN-junction.

PROBLEM 4-8

(a) Show that Eqs. (4-13) and (4-26) can be combined to give

$$u_0 = \frac{C_{1N} - C_{1P}}{a_N - a_P} \tag{4-27}$$

where

$$C_{1N} = u_{mN} \sinh u_{mN} - \cosh u_{mN} \tag{4-28}$$

$$a_N = \sinh u_{mN} \tag{4-29}$$

(b) Show that for the junction of Fig. 4-3

$$u_0 = -9.8 \tag{4-30}$$

(c) Verify that the shape of the junction in the N-region of Fig. 4-3 is given by

$$r = \int_{-9.8}^{u} \frac{du}{\sqrt{\dfrac{\cosh u}{23.5} - u + 2.85}} \tag{4-31}$$

using the method of Problem 4-6.

By combining the results of the above problem with a similar numerical integration for the P-region, we obtain the complete energy band diagram for a germanium junction with 10^{15} donors/cm^3 and 10^{18} acceptors/cm^3. This method can be used for any selected set of impurities.

Figure 3-14 indicates that the PN-junction is equivalent to a capacitor, as discussed in Section 1-4, and the capacity of an abrupt junction can be estimated by solving Poisson's equation using some simplifying assumptions, following Muss (1955)[4]. Let the limits of the space charge layer be located at $x = x_P$ and $x = x_N$, and the P region of the space charge layer have a uniformly-ionized acceptor density N_A right up to the edge of the space-charge region, rather than the charge distribution determined in Problem 4-2c. Similarly, let the N region have a uniformly-ionized donor density, N_D, giving the following solution of Eq. (4-8):

$$\mathcal{V} = (x - x_P)^2 N_A e/2\epsilon + C_1, \qquad x < 0$$

$$\mathcal{V} = -(x - x_N)^2 N_D e/2\epsilon + C_2, \qquad x > 0$$

PROBLEM 4-9

(a) Verify the solution of Eq. (4-8) given above.

(b) If $\mathcal{E} = -d\mathcal{V}/dx$ is continuous at $x = 0$, show that x_P and x_N can be expressed in terms of the width W of the space-charge layer by

$$x_P = -N_D W/(N_D + N_A)$$

$$x_N = N_A W/(N_D + N_A)$$

(c) Show that if we choose $x = 0$ as the point where $\mathcal{V} = 0$, in order to evaluate C_1 and C_2, and if the built-in voltage $\mathcal{V}_0$ is expressed by

$$\mathcal{V}_0 = -\mathcal{V}(x_P) + \mathcal{V}(x_N)$$

then

$$\mathcal{V}_0 = \frac{e\,N_D\,N_A\,W^2}{2\epsilon(N_D + N_A)}$$

(d) Using the definitions given by Eqs. (1-16) and (1-18), and the results of Problem 4-1, show that the capacity C_A per unit area of the junction can be expressed

$$C_A = \frac{q_A}{\mathcal{V}} = \epsilon/W$$

(e) Show that

$$C_A^{-2} = (2\mathcal{V}_0/\epsilon e)\left(\frac{1}{N_D} + \frac{1}{N_A}\right)$$

so that C_A is proportional to $1/\mathcal{V}_0^2$.

4-2 The Low-Level Junction

Section 4-1 shows how to calculate the energy-band diagram for the PN-junction in equilibrium. We are now ready to consider the corresponding problem for a junction to which a voltage is applied, as shown in Fig. 4-4. Before doing this, however, let us again look at the situation in an intuitive way.

If the applied voltage starts at zero and is increased so that the P-region is positive, then the voltage opposes the barrier field $\mathcal{E}$ of Fig. 3-14 and eventually this field will be neutralized so that electrons will diffuse into the P-region and holes into the N-region. With reference to Fig. 3-15, the field $\mathcal{E}$ corresponds to the barrier potential $\mathcal{V}_0$, which for the symmetric junction with 10^{15} donors/cm^3 and 10^{15} acceptors/cm^3, is about 0.2 volt.

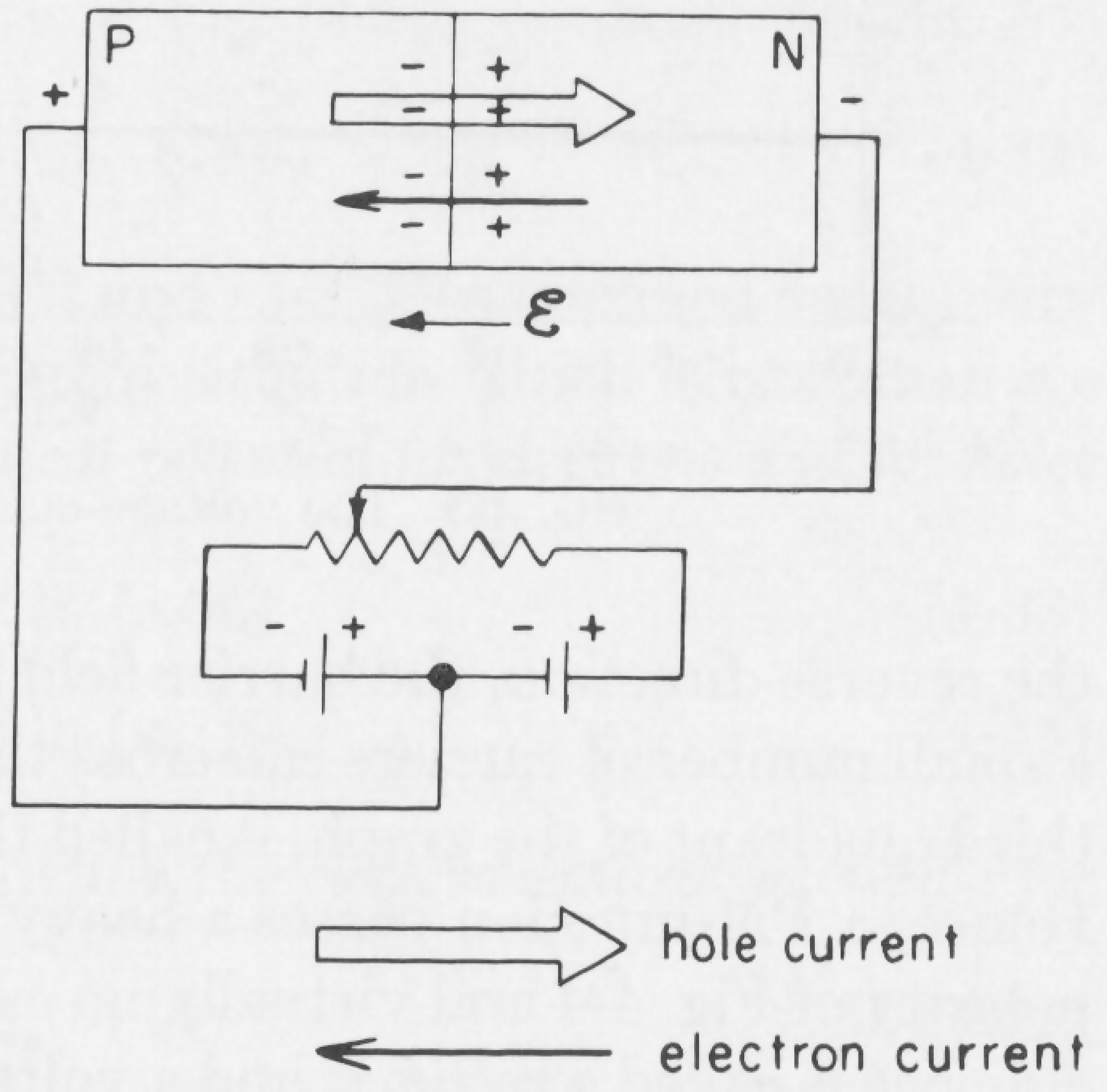


FIG. 4-4. The effect of an applied voltage on the space charge layer in a junction.

PROBLEM 4-10 What is $\mathcal{V}_0$ for the junction of Fig. 4-3?

As the applied voltage reaches and then exceeds $\mathcal{V}_0$, the flow of electrons and holes (in opposite directions) takes a sharp rise and the current-voltage curve looks like the first quadrant of Fig. 4-5. We should emphasize at this point that for low voltages, diffusion is the main cause of carrier flow outside the space-charge region; the contribution due to the applied field, called the *field* or *drift* current, becomes comparable to the diffusion current only for high fields. If we now return the voltage to zero and build it up in the reverse direction, the barrier field is aided by the applied field, and only a small number of carriers can cross the barrier. This current, shown in the third quadrant of the graph, is called the *leakage* or *reverse saturation current.* Hence, a PN-junction passes a heavy current for an applied voltage of the polarity of Fig. 4-4 and virtually no current for the opposite polarity. Such a device is called a *rectifier*, and a voltage with the polarity shown is called a *forward bias voltage.* (A useful mnemonic device is that forward bias corresponds to a Positive P-region.) From the description of the junction fabrication given in Section 3-4, we realize that the P-region is more heavily doped than the N-region, and we therefore expect the hole current from P-region to N-region to be considerably greater than the electron current in the other direction. The introduction of this substantial hole current into the N-region is called *hole injection.* The injected hole current diffuses towards the negative contact, whose bias is of the appropriate polarity to encourage this process. Some of the holes may *recombine* with electrons in the N-region, and this recombination process reduces the efficiency of the junction. The length of time for which a hole is likely to exist without recombining in the presence of a larger number of electrons is called the *lifetime* τ. Since electrons are the majority carrier in N-type germanium, τ is also called the *minority carrier lifetime*, and values of τ of the order of

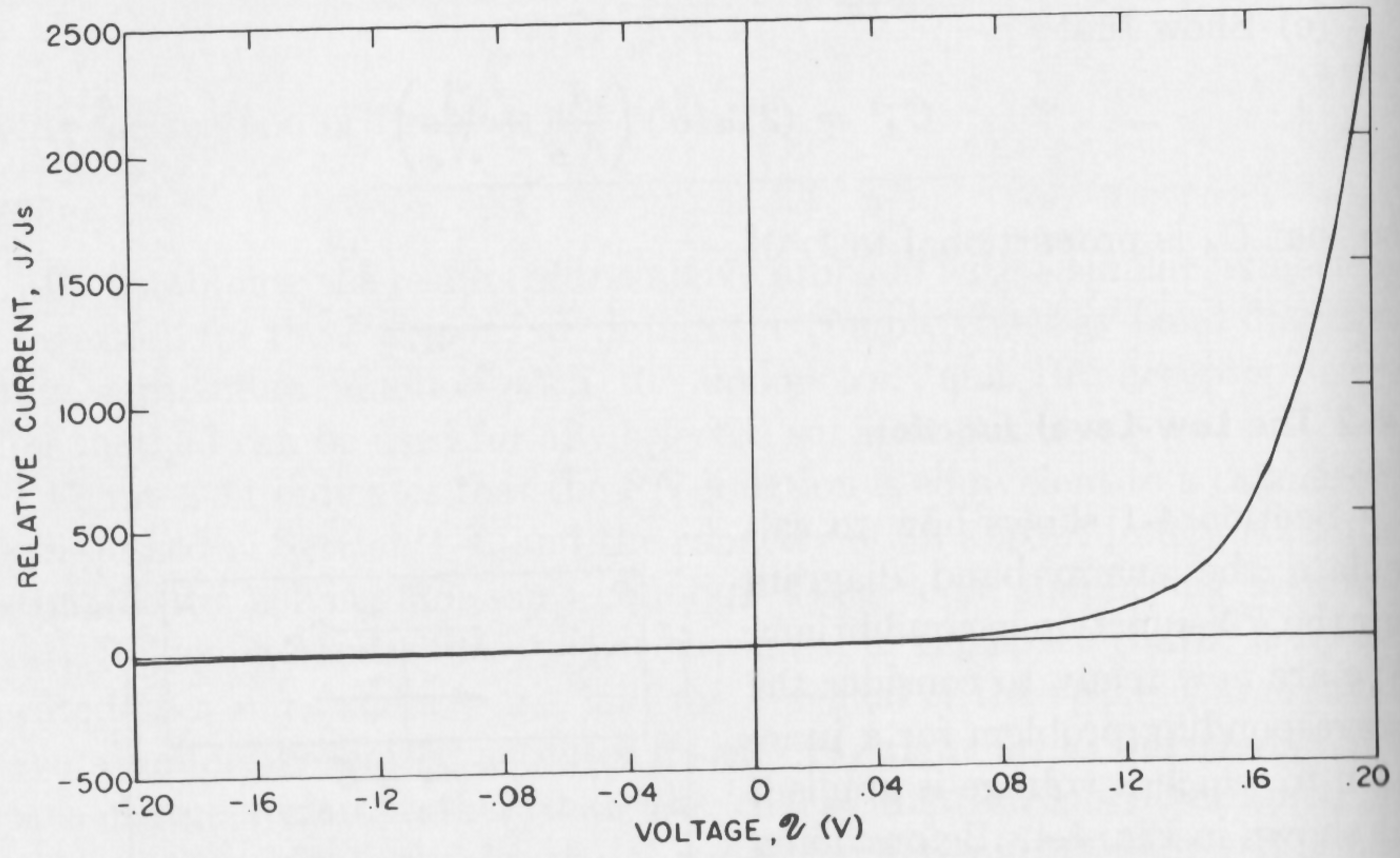


FIG. 4-5. The voltage-current curve of a junction.

100 μsec indicate good quality starting material. We shall discuss the topic of minority carrier lifetime in a more precise fashion later in this chapter.

When a forward voltage $\mathcal{V}$ is applied across the PN-junction of Fig. 4-3, holes will be injected into the N-region and electrons into the P-region, as described above. For a device like this one, the equilibrium concentration of holes p_{0P} in the P-region is considerably greater than the equilibrium concentration of electrons n_{0N} in the N-region, so we would expect that the bulk of the injected current will consist of holes going into the N-region. These holes tend to build up a large electric field in the N-region which adds to the barrier field already existing at the junction. However, this new field will be self-cancelling, since electrons will be urged in from the battery, and this effect is known as *space-charge neutralization.* The concentration of injected holes, denoted by Δp, is thus balanced by an increase Δn in electron concentration, and both n_{0N} and p_{0N} *increase.* A similar effect occurs in the P-region, so everywhere in the junction, Eq. (3-24) becomes modified to

$$np = (n_0 + \Delta n)\,(p_0 + \Delta p) > n_i^2 \qquad (4\text{-}32)$$

The chemical potential K of Eqs. (3-32) and (3-33) was defined for equilibrium conditions. For the non-equilibrium situation which exists when we have injection, we must define a *chemical potential for electrons and for holes* separately by

$$K_n = kT \ln (n/n_i) \qquad (4\text{-}33)$$

$$K_p = kT \ln (n_i/p) \qquad (4\text{-}34)$$

In terms of K_n and K_p, we can then define the *electrochemical potentials for electrons and holes* as

$$\bar{\mu}_n = K_n - e\mathcal{V} \qquad (4\text{-}35)$$

$$\bar{\mu}_p = K_p - e\mathcal{V} \qquad (4\text{-}36)$$

Subtracting Eq. (4-36) from Eq. (4-35) gives

$$\bar{\mu}_n - \bar{\mu}_p = kT \ln (np/n_i^2)$$

or

$$np = n_i^2\, e^{(\bar{\mu}_n - \bar{\mu}_p)/kT} \qquad (4\text{-}37)$$

This reduces to

$$np = n_i^2 \qquad (3\text{-}20)$$

only when $\bar{\mu}_n = \bar{\mu}_p$, so for regions where injection is occurring, the electrochemical potentials for electrons and holes are distinct. This separation of $\bar{\mu}_n$ and $\bar{\mu}_p$ at the junction region is shown in Fig. 4-6a. In this figure, the position of $\bar{\mu}$ at the left-hand end of the P-region is lower than the position at the right-hand end of the N-region; the reason for this will be investigated

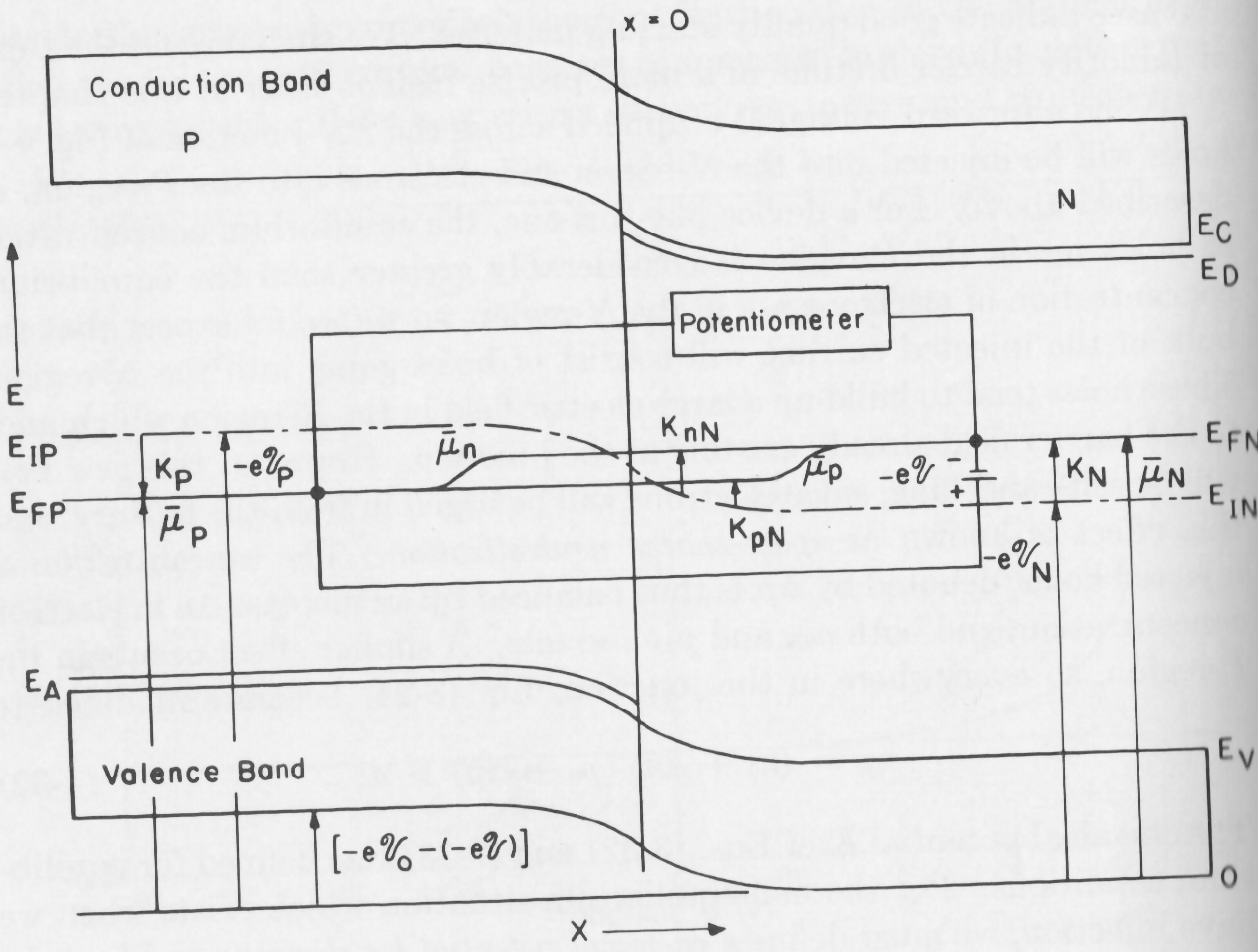


FIG. 4-6(a). Energy level diagram for a PN-junction with an applied forward voltage.

below. We remind the reader at this point that the concept of Fermi level is not defined in a non-equilibrium situation, and we can no longer identify E_F and $\bar{\mu}$. To emphasize this, some writers, such as Shockley (1950)[5], refer to $\bar{\mu}_n$ and $\bar{\mu}_p$ as *quasi-Fermi levels.*

Let us trace the behavior of $\bar{\mu}_p$ as we start at the far left of the P-region and travel the length of the junction. At the starting point, the concentrations will be n_0 and p_0, so that $\bar{\mu}_{nP} = \bar{\mu}_{pP}$. As we move towards the junction, we begin to see the injected electron current, and n_0 is increased by Δn. The hole concentration also goes up by Δp, for a reason to be considered in Problem 4-11 but in a highly doped P-type region, the original concentration p_0 will not be noticeably affected. As n increases, Eq. (4-33) shows that K_n follows suit, and, by Eq. (4-35), $\bar{\mu}_n$ does also. Thus, as the figure shows, $\bar{\mu}_p$ holds steady until we get through the space-charge layer, but $\bar{\mu}_n$ splits off and rises. As we approach from the far right, there is a similar splitting, with $\bar{\mu}_n$ and $\bar{\mu}_p$ interchanging roles.

Now let us attempt to measure the applied voltage $\mathcal{V}$ using an instrument which draws no current. Following Domenicali (1954)[6], Eq. (2-102) shows that when $J = 0$ in the potentiometer circuit, then $d\bar{\mu}/dx = 0$, or $\bar{\mu}$ is constant. Just on either side of the boundary between the probe and the

---

P-region, the value of $\bar{\mu}$ must then be $\bar{\mu}_p$, and this is true up to the potentiometer terminal. Similarly, for the other probe and the other potentiometer terminal, $\bar{\mu} = \bar{\mu}_N$, so the potentiometer measures $\bar{\mu}_N - \bar{\mu}_P$. If the two leads to the potentiometer are the same material — copper, for example — then the chemical potential of the electrons will be the same at both terminals. Hence

$$\bar{\mu}_N - \bar{\mu}_P = e(\mathcal{V}_P - \mathcal{V}_N) = -e\mathcal{V}$$

where $-e\mathcal{V}$ is the energy corresponding to the applied forward voltage. Thus, the potentiometer measures the difference in equilibrium electrochemical potentials or Fermi levels between the P- and N-regions, and this difference is equal to the applied voltage multiplied by $-e$. This is shown in Fig. 4-6a, which has been drawn using the convention that the N-region

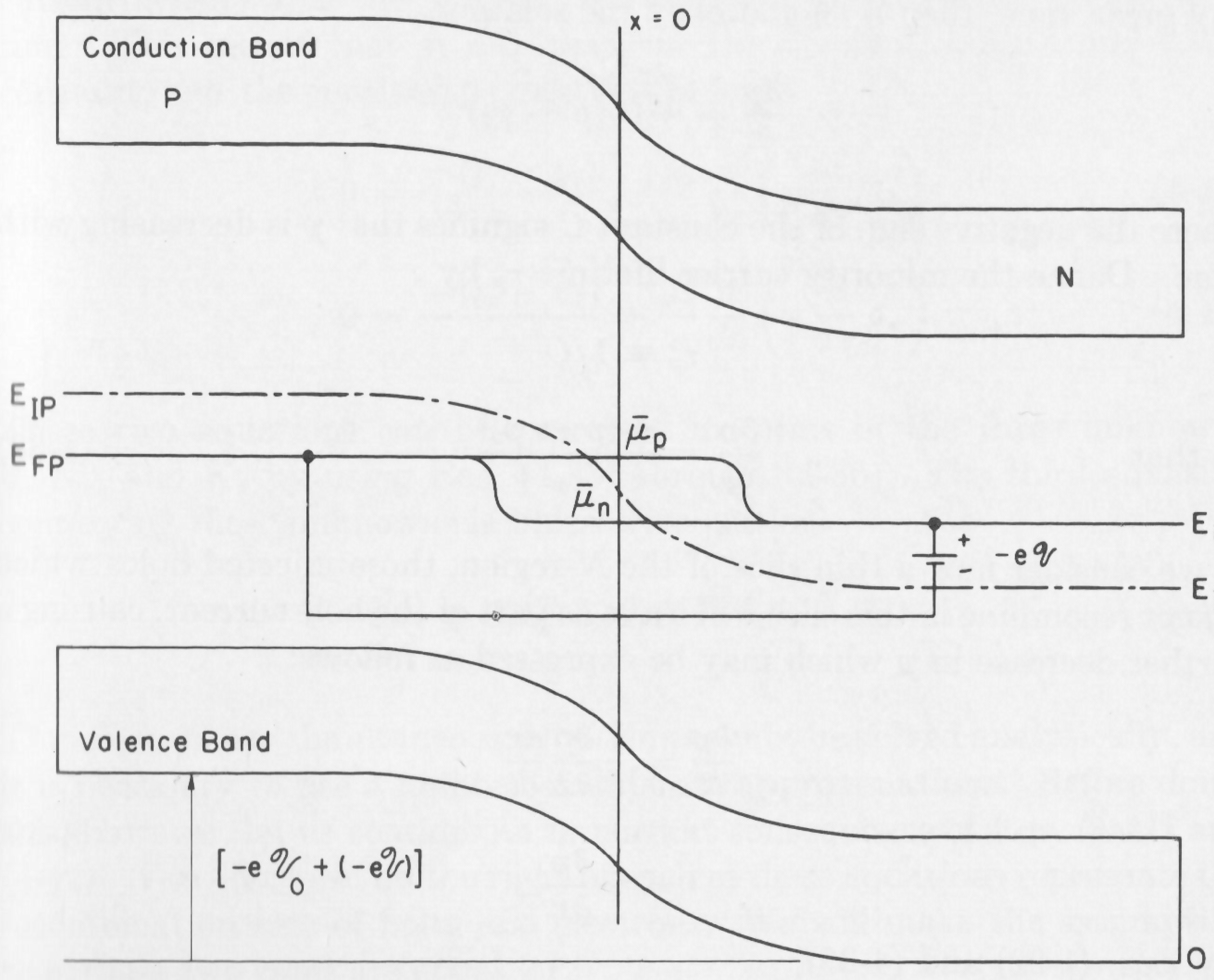


FIG. 4-6(b). Energy level diagram for a PN-junction with an applied reverse voltage.

remains at a fixed potential and the applied potential causes the energy levels in the P-region to shift. Note that $\mathcal{V}$ has the effect of decreasing the barrier $-e\mathcal{V}_0$ to $(-e\mathcal{V}_0 + e\mathcal{V})$ and thus increasing the diffusion currents. When a reverse bias is applied, as in Fig. 4-6b, the Fermi levels or equilibrium electrochemical potentials $\bar{\mu}_N$ and $\bar{\mu}_P$ shift in the opposite direction and the barrier is raised to $-e(\mathcal{V}_0 + \mathcal{V})$. Since there is very little current flowing, non-equilibrium conditions exist only in a region close to the junction, and the splitting of the $\bar{\mu}$'s occurs only in the space-charge region. In the figure, we have made the fundamental assumption that both the N-type and P-type materials have a resistivity which is low enough to justify the assumption that all the applied voltage appears across the junction.

With this descriptive introduction to the non-equilibrium PN-junction, we can now consider quantitatively the behavior of a junction for conditions of low-level injection. When holes are injected into the N-region of a PN-junction, the hole density rises from p_0 to p, and it then decays back to p_0 due to recombination with electrons. If we assume that this process takes place at a rate proportional to the number of excess holes present at any given time, then it should obey the relation

$$\frac{\partial p}{\partial t} = -C\,(p - p_0)$$

where the negative sign of the constant C signifies that p is decreasing with time. Define the minority carrier lifetime τ_p by

$$\tau_p \equiv 1/C$$

so that

$$\frac{\partial p}{\partial t} = \frac{-(p - p_0)}{\tau_p} \tag{4-38}$$

If we consider now a thin slice of the N-region, those injected holes which do not recombine in this slice will leave as part of the hole current, causing a further decrease in p which may be expressed as follows:

$$\begin{aligned} \frac{\partial p}{\partial t} &= -\frac{\partial p}{\partial x}\frac{dx}{dt} \\ &= -\frac{\partial p}{\partial x}v \end{aligned} \tag{4-39}$$

By Eqs. (1-22) and (4-39),

$$\frac{\partial J_p}{\partial x} = \frac{\partial p}{\partial x}ev = -e\frac{\partial p}{\partial t} \tag{4-40}$$

If we combine Eqs. (4-38) and (4-40), the total rate of decrease of holes at any part of the N-region is

$$\frac{\partial p}{\partial t} = \frac{-(p - p_0)}{\tau_p} - \frac{1}{e}\frac{\partial J_p}{\partial x} \tag{4-41}$$

This is called the *equation of continuity* for holes. For electrons, the corresponding equation is

$$\frac{\partial n}{\partial t} = \frac{-(n - n_0)}{\tau_n} + \frac{1}{e}\frac{\partial J_n}{\partial x} \tag{4-42}$$

The generalized Ohm's law, Eq. (2-102), enables us to express the currents as

$$J_n = (\sigma/e)\frac{d\,\bar{\mu}_n}{dx} = n\,\mu_n\frac{d\,\bar{\mu}_n}{dx} \tag{4-43}$$

$$J_p = p\,\mu_p\frac{d\,\bar{\mu}_p}{dx} \tag{4-44}$$

We will now assume what are called *steady-state conditions;* that is, the total current density J in the junction is a constant and does not change with time. This means that at any point in the junction, both n and p are constant, and the continuity equations become

$$0 = \frac{-(p - p_0)}{\tau_p} - \frac{\mu_p}{e}\frac{d}{dx}\left(p\frac{d}{dx}\bar{\mu}_p\right) \tag{4-45}$$

$$0 = \frac{-(n - n_0)}{\tau_n} + \frac{\mu_n}{e}\frac{d}{dn}\left(n\frac{d}{dx}\bar{\mu}_n\right) \tag{4-46}$$

These two equations can be expressed in terms of the three unknowns $\mathcal{V}$, K_n, and K_p by using Eqs. (4-33) through (4-36). The third equation connecting these unknowns is Poisson's equation.

$$\frac{d^2\mathcal{V}}{dx^2} = \frac{-(p - n + N_D - N_A)}{\epsilon} \tag{4-47}$$

These non-linear simultaneous equations cannot be solved analytically, and it is necessary to use a number of different approximations. Before doing this, however, let us consider an important consequence of Eqs. (4-41) and (4-42). The first term on the right of each of these equations represents the recombination rate of holes and electrons. We will make the assumption that these two rates are equal, or

$$\frac{-(p - p_0)}{\tau_p} = \frac{-(n - n_0)}{\tau_n} \tag{4-48}$$

This is reasonable in view of the fact that the greater the injected concentration, the longer the time it takes for the excess carriers to decay (later in this chapter we shall show how lattice imperfections affect this assumption).

PROBLEM 4-11

(a) Show that Eqs. (4-40), (4-41), and (4-48) can be combined to give

$$\frac{dJ}{dx} - \frac{\partial q_V}{\partial t} = 0 \tag{4-49}$$

where $J = J_n + J_p$

(b) Show how Eq. (4-49) can be converted into a differential equation whose solution is

$$q_V = q_{V0}\, e^{-t/\tau} \tag{4-50}$$

What do the symbols q_{V0} and τ represent? Show that τ for a typical semiconductor is of the order of micromicroseconds.

The above problem indicates that excess charge in a semiconductor decays very rapidly and the junction, except for the narrow space-charge region, contains no net charge. This idea is known as the hypothesis of *space-charge neutrality*.

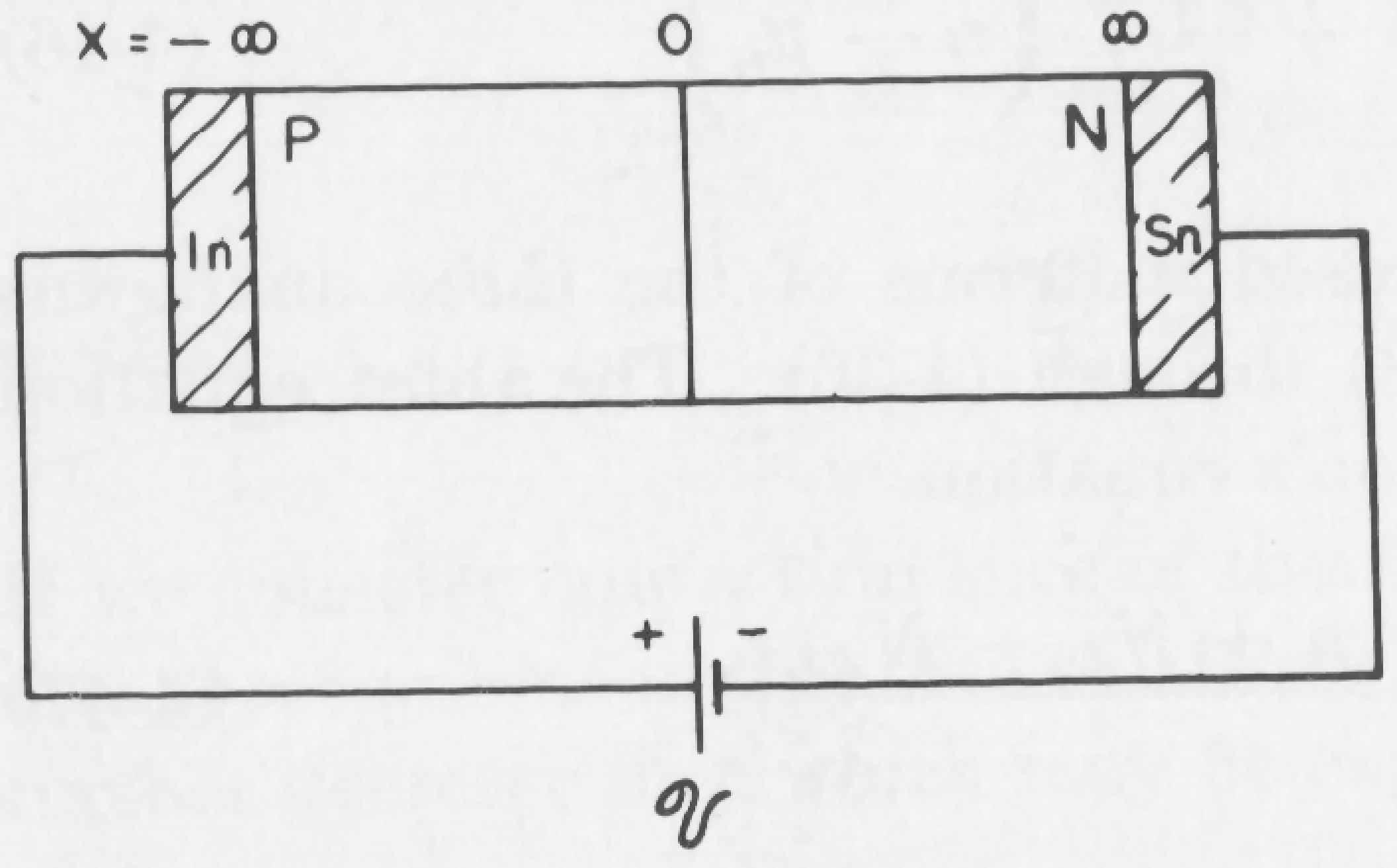


FIG. 4-7. The *PN*-junction with non-rectifying contacts.

Let us now consider the junction of Fig. 4-7. We shall regard the space-charge layer as being so narrow that it is concentrated at $x = 0$, and the boundary between the *P*-region and the indium button is so far from the junction that its position can be designated by $x = -\infty$. The contact to the *N*-region, at $x = \infty$, is made by alloying a tin button, for a reason to be given below. We will assume that the applied forward voltage $\mathcal{V}$ is low enough so that there is no electric field $\mathcal{E}$ in the diode except for the barrier field right at the junction, and this field is not considered in the analysis given here. The hole current J_p in the base, by the equation corresponding to Eq. (3-37), is

$$J_p = -e\, D_p \frac{dp}{dx} \tag{4-51}$$

Equation (4-41) for steady-state conditions is then

$$\frac{d^2 p_N}{dx^2} - \frac{(p_N - p_{0N})}{L_p^2} = 0 \tag{4-52}$$

where p_N is the hole concentration in the *N*-region during injection, p_{0N} is the original concentration, and where we define the *diffusion length* for holes L_p by

$$L_p = (D_p \tau_p)^{1/2} \tag{4-53}$$

This quantity may be regarded as the average distance which a hole will diffuse before recombining.

PROBLEM 4-12 The mobilities of electrons and holes in germanium and silicon are given by Conwell (1958)[2] as shown in Table P4-1.

TABLE P4-1

	Ge	Si
μ_n	3900 cm²/v-sec	1350 cm²/v-sec
μ_p	1900	480

Calculate L_n and L_p for $\tau = 10$, 100, and 1000 μsec in each element.

The solution of Eq. (4-52) is

$$p_N - p_{0N} = Ae^{-x/L_p} + Be^{x/L_p} \tag{4-54}$$

The boundary condition at $x = 0$ comes from Eqs. (4-32) and (4-37).

$$p_N(n_{0N} + \Delta n) = p_{0N}\, n_{0N}\, e^{(\bar{\mu}_n - \bar{\mu}_p)/kT} \tag{4-55}$$

If space-charge neutrality is maintained, then $\Delta n = \Delta p$, and for low-level injection $\Delta p \ll n_{0N}$. Hence, Eq. (4-55) reduces to

$$p_N = p_{0N}\, e^{(\bar{\mu}_n - \bar{\mu}_p)/kT} \tag{4-56}$$

It is seen from Fig. 4-6a that $(\bar{\mu}_n - \bar{\mu}_p)$, which is measured at $x = 0$ in Eq. (4-55), equals $+ e\mathcal{V}$, so

$$p_N = p_{0N}\, e^{e\mathcal{V}/kT} \tag{4-57}$$

At $x = \infty$, there is a tin contact, which has been found experimentally to be an *infinite recombination rate* contact; that is, all holes which succeed in reaching it instantly recombine. Hence, the boundary condition at $x = \infty$ is

$$p_N = p_{0N} \tag{4-58}$$

PROBLEM 4-13

(a) Show that the solution of Eq. (4-54), subject to the boundary conditions, is

$$p_N - p_{ON} = p_{ON}(e^{e\mathcal{V}/kT} - 1)e^{-x/L_p} \tag{4-59}$$

(b) Show that the corresponding solution for the P-region is

$$n_P - n_{OP} = n_{OP}(e^{e\mathcal{V}/kT} - 1)e^{x/L_n} \tag{4-60}$$

Substituting Eq. (4-59) into Eq. (4-34) yields

$$K_{pN} = kT \ln \{n_i/[p_{ON} + p_{ON}(e^{e\mathcal{V}/kT} - 1)e^{-x/L_p}]\} \tag{4-61}$$

As $x \to L_p$, the term involving e^{-x/L_P} becomes small and K_{pN} approaches its equilibrium value K_N, so that $\bar{\mu}_p$ rises to merge with the common $\bar{\mu}_N$ line, confirming Fig. 4-6a. A similar behavior is shown by K_{nP}, but the curves are not symmetrical for a typical junction owing to the large difference in impurity concentrations of the two regions.

PROBLEM 4-14

(a) Show that the hole current at the junction is

$$J_p = (e\,D_p\,p_{ON}/L_p)(e^{e\mathcal{V}/kT} - 1) \tag{4-62}$$

(b) Show that the electron current at the junction is

$$J_n = (e\,D_n\,n_{OP}/L_n)(e^{e\mathcal{V}/kT} - 1) \tag{4-63}$$

When the junction is reverse-biased, then $e^{e\mathcal{V}/kT}$ will vanish for large voltages and the *reverse saturation* or *leakage current* J_s is

$$J_s = J_{ps} + J_{ns} = (e\,D_p\,p_{ON}/L_p) + (e\,D_n\,n_{OP}/L_n) \tag{4-64}$$

Hence, the junction is carrying a current given by

$$J = J_s(e^{e\mathcal{V}/kT} - 1) \tag{4-65}$$

A plot of Eq. (4-65) is given in Fig. 4-4, confirming again our intuitive ideas.

We define the *injection efficiency* γ of a PN-junction as the fraction of current through the junction carried by holes, or

$$\gamma = \frac{\partial J_p}{\partial(J_p + J_n)} = \frac{\partial J_p}{\partial J} \tag{4-66}$$

PROBLEM 4-15

(a) Show that the injection efficiency can be expressed by

$$\gamma = \frac{1}{1 + (D_n/D_p)(n_{OP}/p_{ON})(L_p/L_n)} \tag{4-67}$$

Hint: Use $\gamma = \dfrac{\partial J_p}{\partial \mathcal{V}} \Big/ \dfrac{\partial J}{\partial \mathcal{V}}$

(b) Show that γ is approximately 1.0 in the germanium junction of Fig. 4-3.

(c) Show that when $\mathcal{V} \ll 26$ mV, Eq. (4-65) reduces to Ohm's law. What is the significance of this result?

The low-level theory presented here indicates that γ depends only on the mobilities and the impurity concentrations in the two regions of the junction. It will be shown later that at higher injection levels, the diode properties depend strongly on the total current in the junction.

4-3 The Junction Transistor

It is possible to make a semiconductor amplifier by starting with a piece of N-type germanium and adding P-type impurities from each end of the block (Fig. 4-8), getting two PN-junctions, or a *PNP transistor*. The three regions of the transistor are named as shown in the figure, and the base region is made as thin as possible. In Fig. 4-9 we have shown a cross-sectional view of the structure of a PNP alloyed germanium transistor, and in Fig. 4-10 the individual electron and hole currents across the junctions are shown. Note that there is both an electron and a hole current flowing across each junction; since they move in opposite directions, the total current is the numerical sum of the two components. The figure also shows how these internal currents are related to the electron currents carried by the external circuit. Considering the emitter current I_E, for example, we note that this current must equal the

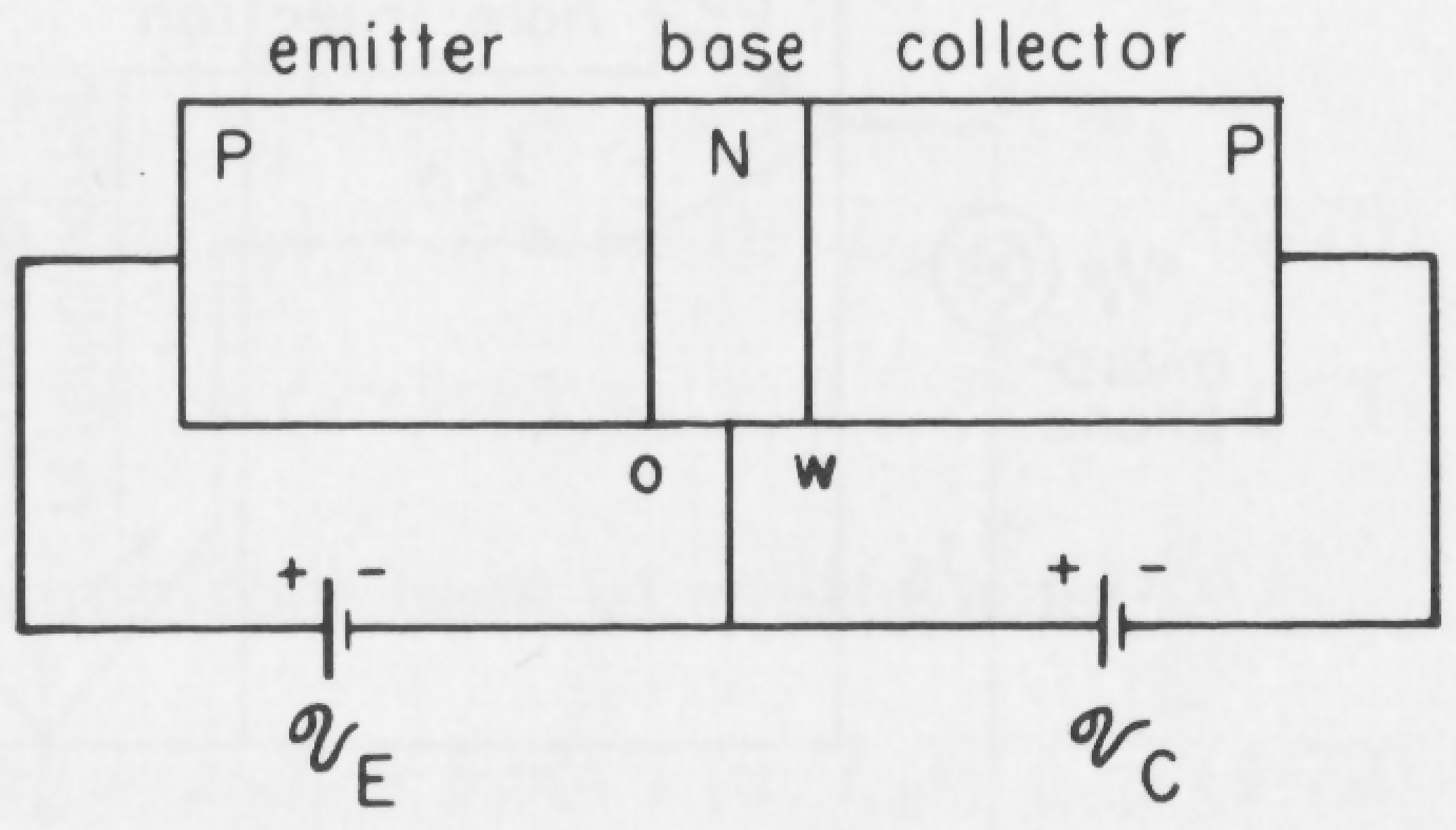


FIG. 4-8. A PNP transistor, showing the names of the three electrodes and the bias arrangement for the common base configuration.

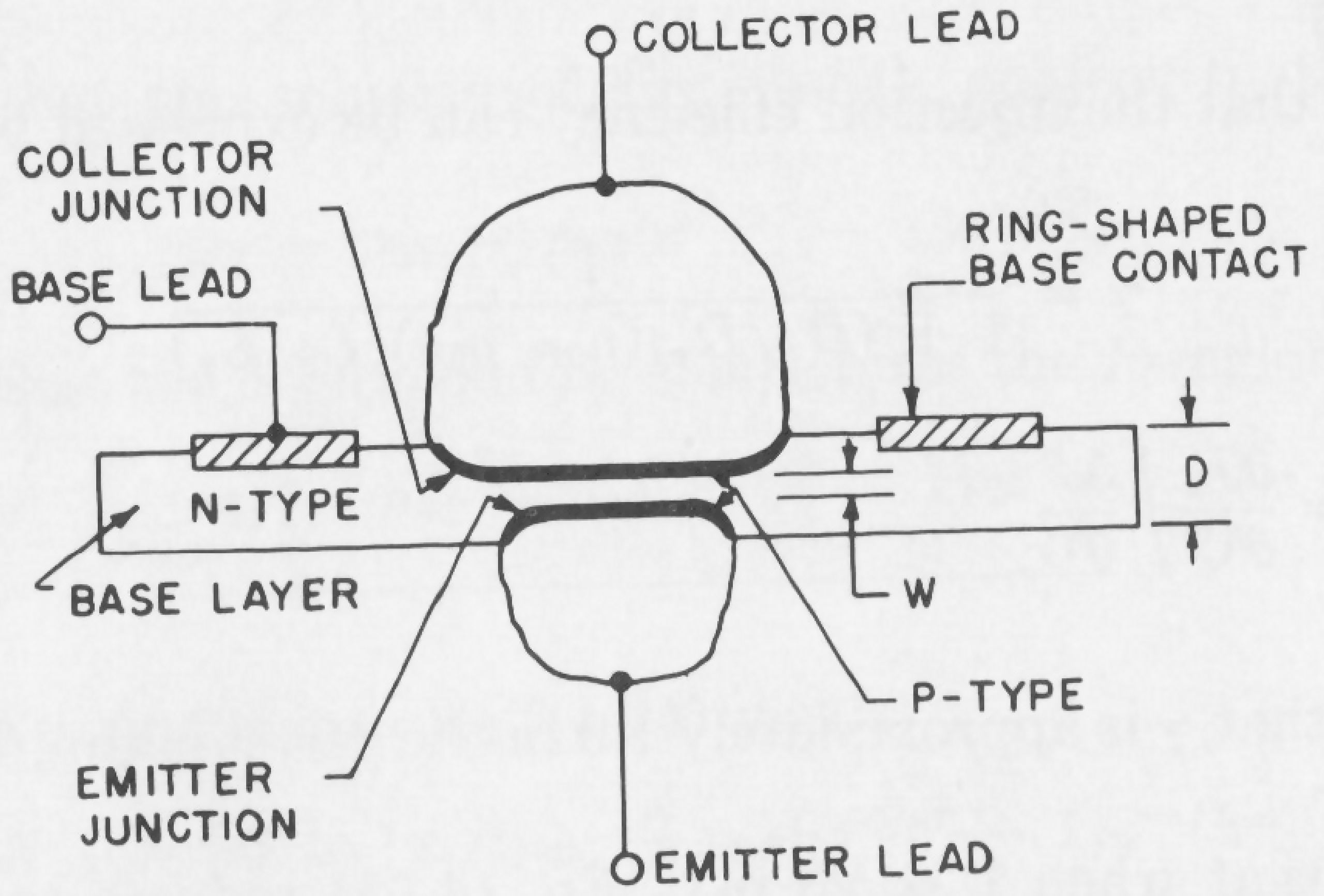

COLLECTOR LEAD
COLLECTOR JUNCTION
BASE LEAD
RING-SHAPED BASE CONTACT
D
N-TYPE
BASE LAYER
W
P-TYPE
EMITTER JUNCTION
EMITTER LEAD

FIG. 4-9. A cross-sectional sketch of a representative *PNP* transistor. (After A. E. Anderson, *The Western Electric Engineer*, July, 1959, with permission.)

sum of its two internal components: the hole current injected into the base and the electron current injected into the emitter. Since hole currents flow only in the transistor, an electron current equal to the injected hole current

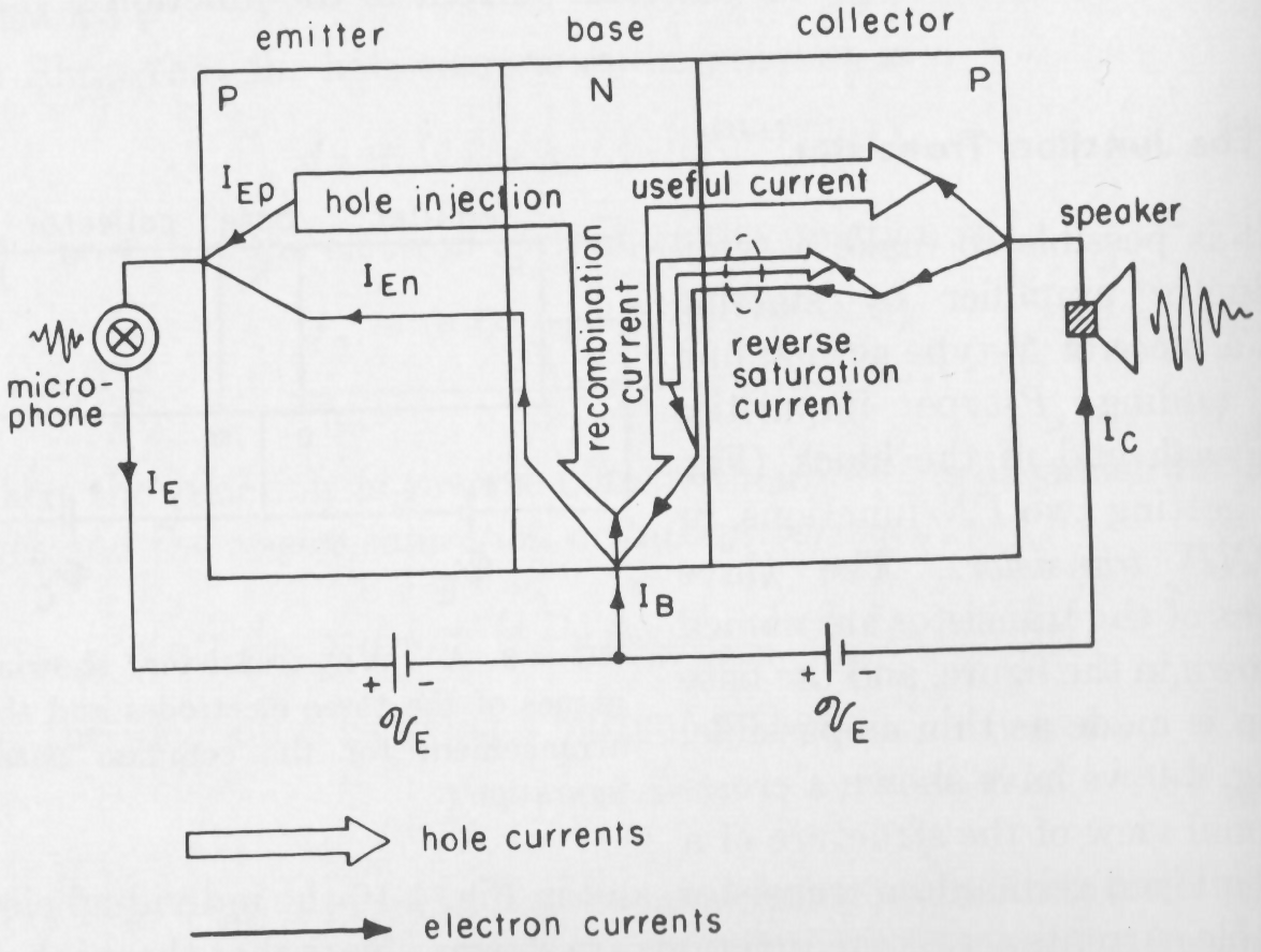

emitter
base
collector
P
N
P
I_{Ep}
hole injection
useful current
speaker
I_{En}
recombination current
reverse saturation current
micro-phone
I_E
I_C
I_B
$\mathcal{V}_E$
$\mathcal{V}_E$
hole currents
electron currents

FIG. 4-10. The internal and external currents for a *PNP* transistor.

shows up externally instead, and this is the current denoted I_{Ep}. Hence, I_{Ep} and I_{En} combine to form I_E. Since the current going into the collector is smaller than the emitter current because of the loss of holes due to recombi-

nation, the transistor has a lower output current than input current. However, the collector junction is reverse-biased, which means that it has a high voltage-to-current ratio, or a high resistance. This indicates that the transistor as shown is carrying considerably more power through the collector junction and the speaker circuit than through the emitter junction and the microphone. Thus, it is acting as a power amplifier. By making the emitter or collector, instead of the base, the common electrode, it is also possible to achieve current or voltage gain in a manner analogous to a vacuum tube.

The transistor of Fig. 4-8, for low injection levels, can be treated by a simple extension of *PN*-junction theory. For the base region we use the solution

$$p_N - p_{0N} = Ae^{-x/L_p} + Be^{+x/L_p} \tag{4-68}$$

subject to the boundary conditions

$$p_N = p_{0N}\, e^{e\mathcal{V}_E/kT} \quad \text{at } x = 0 \tag{4-57}$$

$$p_N = p_{0N}\, e^{+e\mathcal{V}_C/kT} \quad \text{at } x = w \tag{4-69}$$

PROBLEM 4-16

(a) Using Eq. (1-112), show that the hole current injected into the base is

$$J_{pE} = -e\, D_p \left.\frac{dp}{dx}\right)_{x=0}$$

$$= e(p_{0N}\, D_p/L_p)[a \coth (w/L_p) - b \operatorname{csch} (w/L_p)] \tag{4-70}$$

where
$$a \equiv (e^{e\mathcal{V}_E/kT} - 1), \qquad b \equiv (e^{-e\mathcal{V}_C/kT} - 1) \tag{4-71}$$

(b) Show that the hole current going from base to collector is

$$J_{pC} = e(p_{0N}\, D_p/L_p)[a \operatorname{csch} (w/L_p) - b \coth (w/L_p)] \tag{4-72}$$

For the electron currents, the solution corresponding to Eq. (4-68) is

$$n_P - n_{0P} = Ce^{-x/L_n} + De^{x/L_n} \tag{4-73}$$

with boundary conditions

$$n_P = n_{0P}\, e^{e\mathcal{V}_E/kT} \quad \text{at } x = 0; \qquad n_P = n_{0P} \quad \text{at } x = w \tag{4-74}$$

giving
$$J_{nE} = en_{0P}\, D_n\, a/L_n \tag{4-75}$$

and
$$J_{nC} = -en_{0P}\, D_n\, b/L_n \tag{4-76}$$

The total current density through the emitter junction is then

$$\begin{aligned} J_E &= J_{nE} + J_{pE} \\ &= be[(D_n\, n_{0P}/L_e) - (D_p\, p_{0N}/L_p) \coth (w/L_p)] \\ &\qquad + (ae\, D_p\, p_{0N}/L_p) \operatorname{csch} (w/L_p) \end{aligned} \tag{4-77}$$

and for the collector is

$$\begin{aligned} J_C = -be[(D_n\, n_{0P}/L_n) &+ (D_p\, p_{0N}/L_p) \coth (w/L_p)] \\ &+ (ae\, D_p\, p_{0N}/L_p) \operatorname{csch} (w/L_p) \end{aligned} \tag{4-78}$$

Define the *current amplification* α as

$$\alpha = \left.\frac{\partial J_c}{\partial J_E}\right)_{\mathcal{V}_C} \tag{4-79}$$

This can be expressed as

$$\alpha = \left.\frac{\partial J_c}{\partial \mathcal{V}_E}\right)_{\mathcal{V}_C} \Bigg/ \left.\frac{\partial J_E}{\partial \mathcal{V}_E}\right)_{\mathcal{V}_C} \tag{4-80}$$

Substituting Eqs. (4-77) and (4-78) into Eq. (4-80) gives

$$\alpha = \frac{\operatorname{sech} (w/L_p)}{1 + (D_n/D_p)(n_{0P}/p_{0N})(L_p/L_n) \tanh (w/L_p)} \tag{4-81}$$

Using Eq. (4-66) to calculate the injection efficiency γ by the method of Eq. (4-80) yields

$$\gamma = \frac{1}{1 + (D_n/D_p)(n_{0P}/p_{0N})(L_p/L_n) \tanh (w/L_p)} \tag{4-82}$$

Now define the *base transport factor* β as

$$\beta = \left.\frac{\partial J_{pC}}{\partial J_p}\right)_{\mathcal{V}_C} \tag{4-83}$$

and β thus represents the fraction of injected hole current which reaches the collector. Using the same type of calculation as above, we obtain

$$\beta = \operatorname{sech} (w/L_p) \tag{4-84}$$

Hence, Eq. (4-81) may be written

$$\alpha = \beta\gamma \tag{4-85}$$

For a good transistor, $L_p \gg w$, and Eq. (4-82) reduces to an expression almost identical to Eq. (4-67). Further, we still have $n_{0P}/p_{0N} \ll 1$ and hence we expect that $\alpha = \gamma = 1.0$. In practice, we find that this is true only for carefully fabricated units, and then only at low currents. In Fig. 4-11 we show curves [Nussbaum (1955)[7]] for an excellent, an average, and a poor transistor of a group of early power transistors.

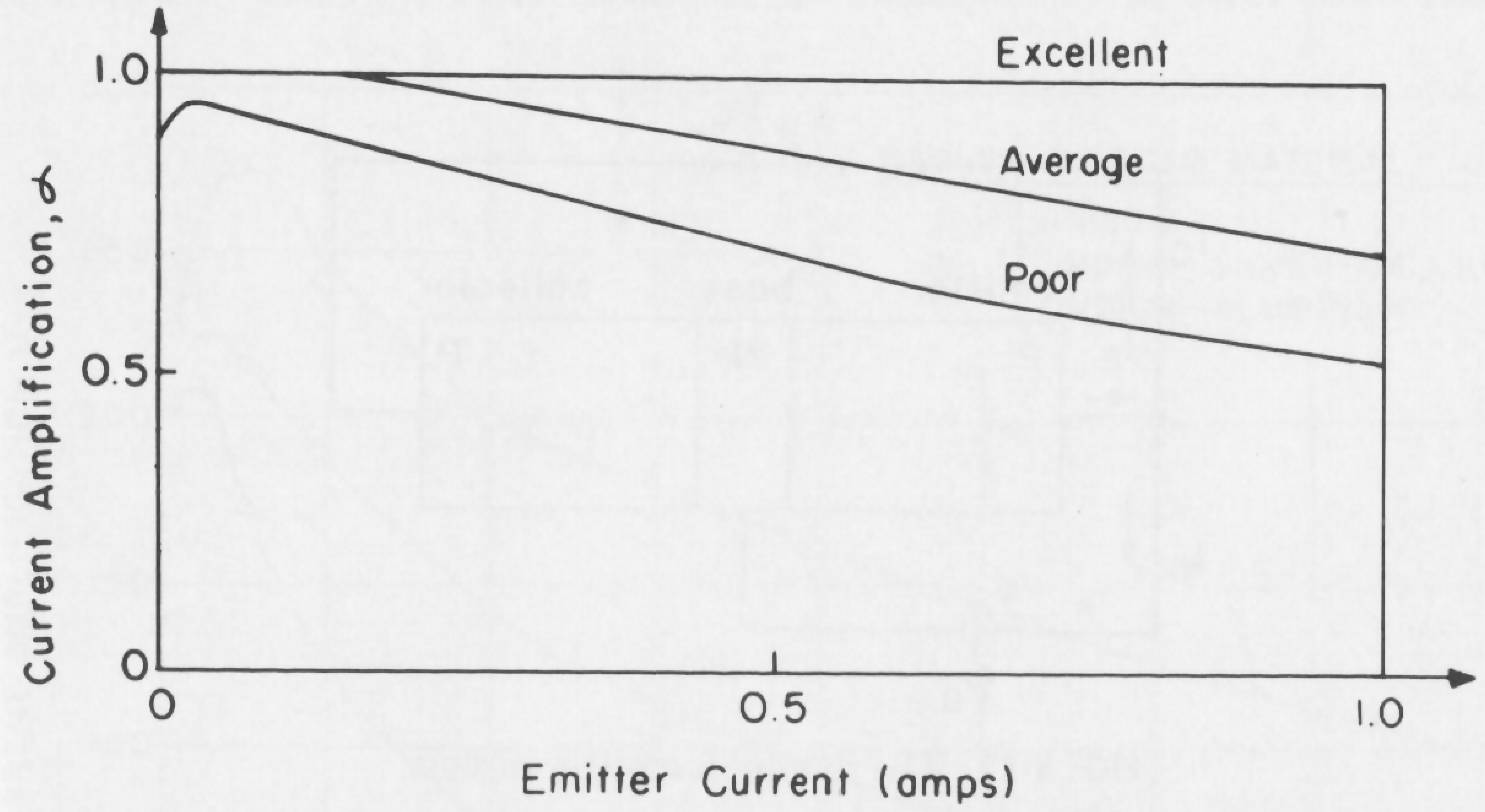


FIG. 4-11. Amplification *vs.* current density for power transistors. (From A. Nussbaum, Proc. IRE **43,** 315, 1955, with permission.)

PROBLEM 4-17

(a) Show that Eqs. (4-77) and (4-78) can be written in the form

$$I_E = G_{EE}B_E + G_{EC}B_C$$

$$I_C = G_{CE}B_E + G_{CC}B_C$$

where I_E is the emitter current, the G's are quantities with the dimensions of ohms^{-1} and are called *conductances,* and B_E and B_C are defined in terms of a and b of Eq. (4-71) as

$$B_E = (kT/e)a, \qquad B_C = (kT/e)b$$

(b) Show that when $I_E = 0$ and $\mathcal{V}_E$ and $\mathcal{V}_C \ll kT/e$

$$\mathcal{V}_E = \alpha\mathcal{V}_C$$

This result provides a way of measuring α at low levels: apply a voltage considerably smaller than 25 mV across the collector junction, leave the emitter open-circuited, and measure $\mathcal{V}_C$.

It was stated above that current amplification could be obtained from a transistor by using the emitter or collector, rather than the base, as the common element. In Fig. 4-12 we show the biasing arrangement for the *common emitter* circuit. In comparing this arrangement with the *common base* circuit of Figs. 4-8 or 4-10, note that once again the emitter junction is

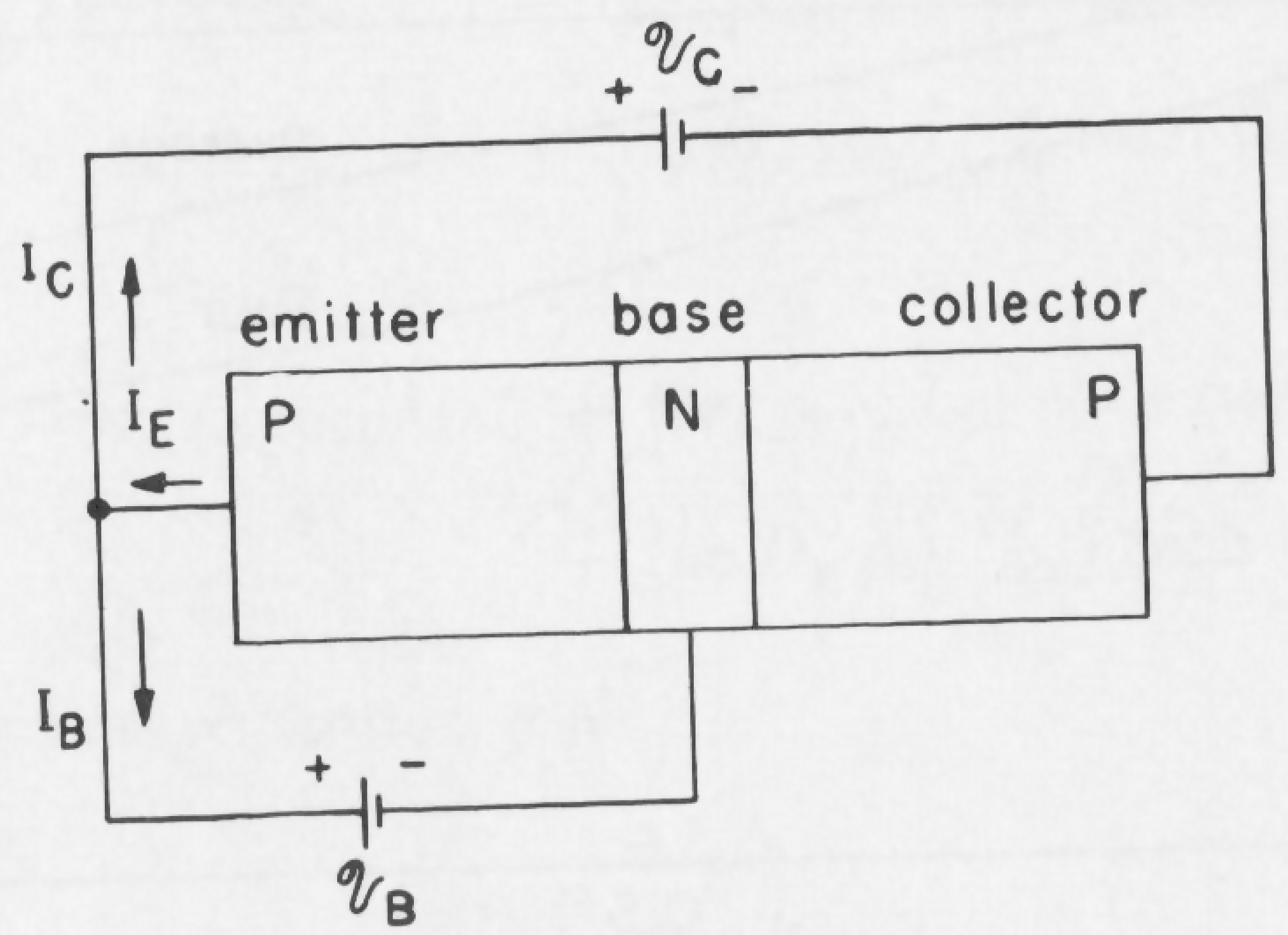


FIG. 4-12. The common emitter circuit.

forward biased and the collector junction is reverse-biased. Note also that for both circuits the emitter, collector, and base currents are related by

$$I_E = I_B + I_C \tag{4-86}$$

For the common emitter circuit, the current amplification α_{CE} is defined, in analogy to Eq. (4-79), as

$$\alpha_{CE} = \left.\frac{\partial J_C}{\partial J_B}\right)_{\mathcal{V}_B} \tag{4-87}$$

PROBLEM 4-18 Denoting the common base current amplification by α_{CB}, show from Eq. (4-86) that

$$\alpha_{CE} = \frac{\alpha_{CB}}{1 - \alpha_{CB}} \tag{4-88}$$

and

$$\alpha_{CB} = \frac{\alpha_{CE}}{1 + \alpha_{CE}} \tag{4-89}$$

(We shall continue to denote α_{CB} by α in cases where no confusion will exist.)

We observe that whereas α_{CB} can have a maximum value of 1.0, α_{CE} is not similarly limited and also, that α_{CE} is a more sensitive indication

of the gain of the transistor. For example, Fig. 4-13 shows the value of α_{CE} for several experimental transistors made by Armstrong, Carlson, and Bentivegna (1956).[8] For curve 3, for example, it is seen that the total

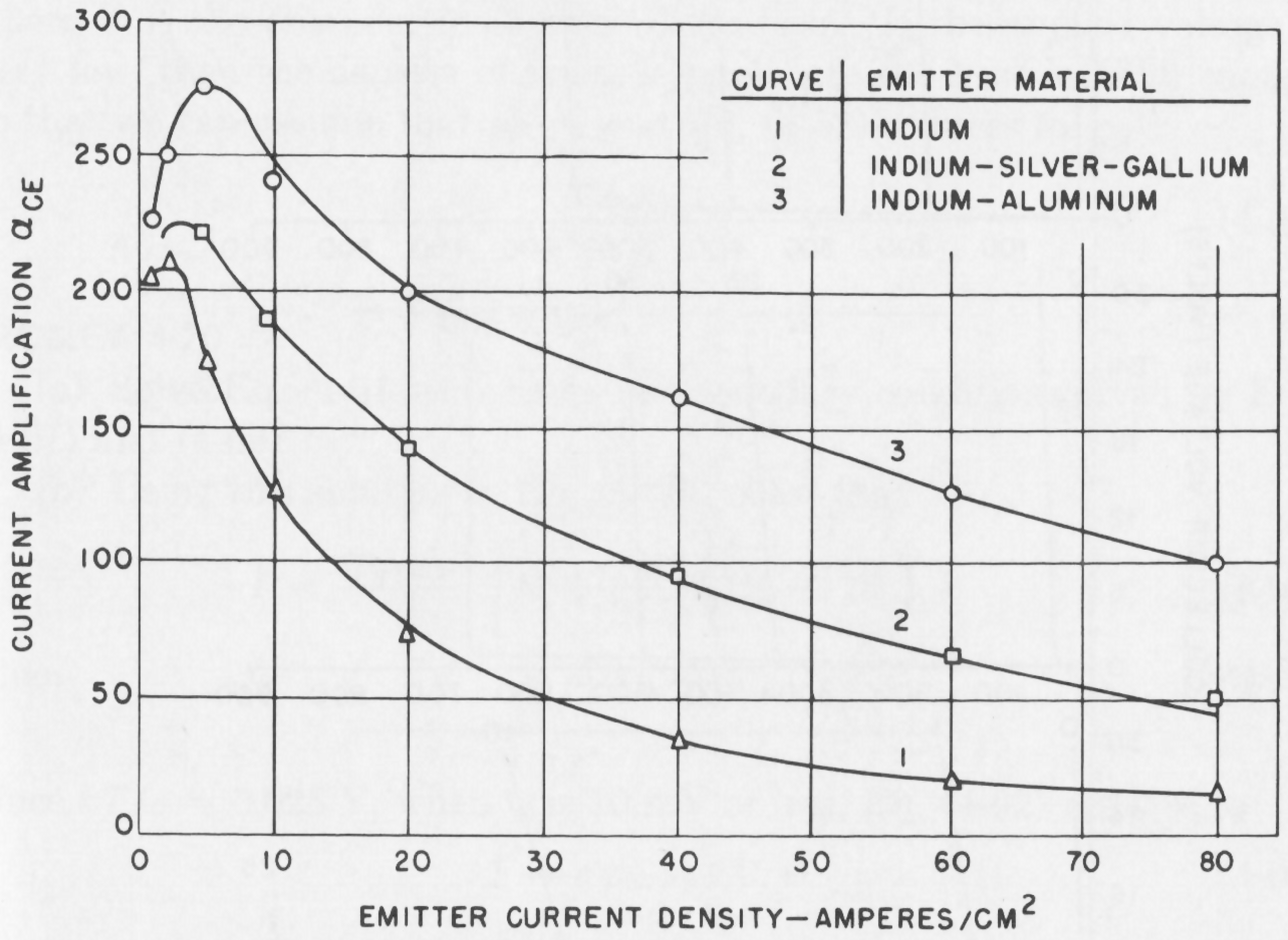


FIG. 4-13. The effect of emitter impurity concentration on efficiency. (From L. D. Armstrong, C. L. Carlson, and M. Bentivegna, RCA Review **17**, 37, 1956, with permission.)

excursion of α_{CE} is 100 to 275, both extremes corresponding to an α_{CB} of 0.99. The figure also indicates the relative efficiency of three different emitter alloying materials, and the high gain of the In-Al alloy can be ascribed in part to the higher doping levels achieved.

A set of common-emitter *characteristic curves* for the three power transistors of Fig. 4-11 is shown in Fig. 4-14. These curves represent the current-voltage relations for the reverse-biased collector junction and in order to be compared with Fig. 4-5, should be rotated 90 deg about the origin in a counter-clockwise direction and then should be reflected in the new position of the voltage axis. It will then be seen that the curves for an open-circuited input ($I_B = 0$) are like the third quadrant of Fig. 4-5, as they should be, although the method of presentation in Fig. 4-14 is the one customarily used. When base current is flowing, as indicated by the figures on the individual curves, we notice that the curves for $I_B = 0$ are shifted to the right. This is because the injected holes in the base pass on through to the collector and add to the leakage current.

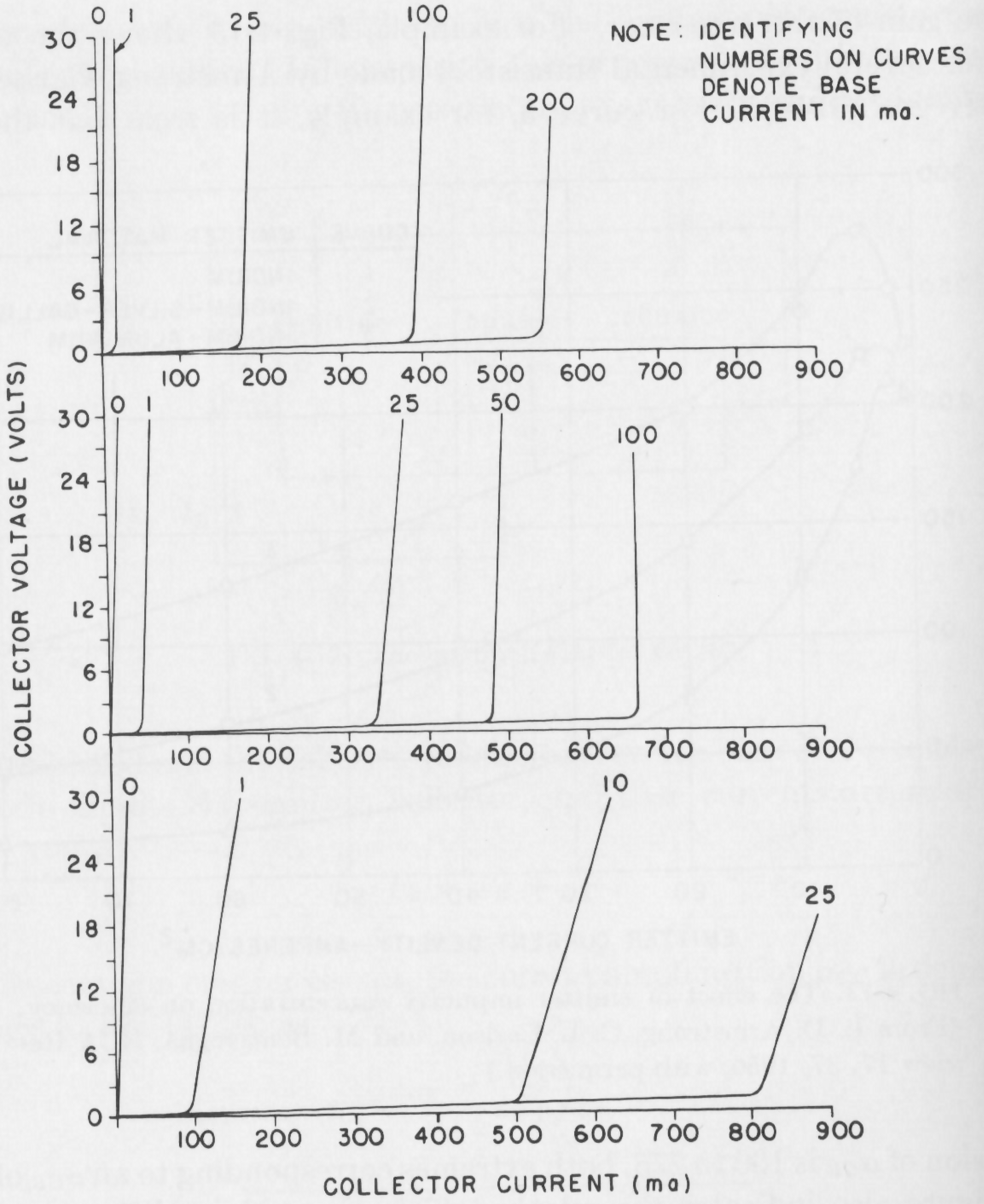


FIG. 4-14. Characteristic curves for the transistors of Figure 4-11. (From Nussbaum, *loc. cit.*)

PROBLEM 4-19 Show qualitatively that the curves of Fig. 4-14 correspond to Eq. (4-78).

As an example of a simple, useful application of the theory given above, consider the problem of determining the basewidth w of a transistor. A method of accomplishing this by purely electrical means has been devised by Long (1957)[9] If we use

$$J_{pE} = -e\,D_p\,\frac{dp}{dx}\Big)_{x=0} \tag{4-70}$$

the current I in the base for a transistor with $\gamma \sim 1.0$ is

$$I = -\mu_p\,kT\,A\,\frac{dp}{dx}\Big)_{x=0} \tag{4-90}$$

where A is the cross-sectional area of the base. If the applied voltage is very low, then the density of holes injected into the base is small enough so that we can assume that $p_N \sim p_{0N}$; Eq. (4-52) reduces to

$$\frac{d^2p_N}{dx^2} = 0 \tag{4-91}$$

PROBLEM 4-20

(a) Solve Eq. (4-91) subject to the boundary conditions given by Eqs. (4-57) and (4-69).

(b) Using this solution in Eq. (4-90), show that

$$I = \frac{ep_{0N}\mu_p A}{w}\left[\mathcal{V} + \frac{e}{2kT}\left(\mathcal{V}_E^2 - \mathcal{V}_C^2\right) + \cdots\right] \tag{4-92}$$

where

$$\mathcal{V} = \mathcal{V}_E + \mathcal{V}_C \tag{4-93}$$

Since $kT/e = 0.026$ V, when $\mathcal{V}$ is 10 mV or less, Eq. (4-92) reduces to

$$I = e\,p_{0N}\mu_p A\mathcal{V}/w \tag{4-94}$$

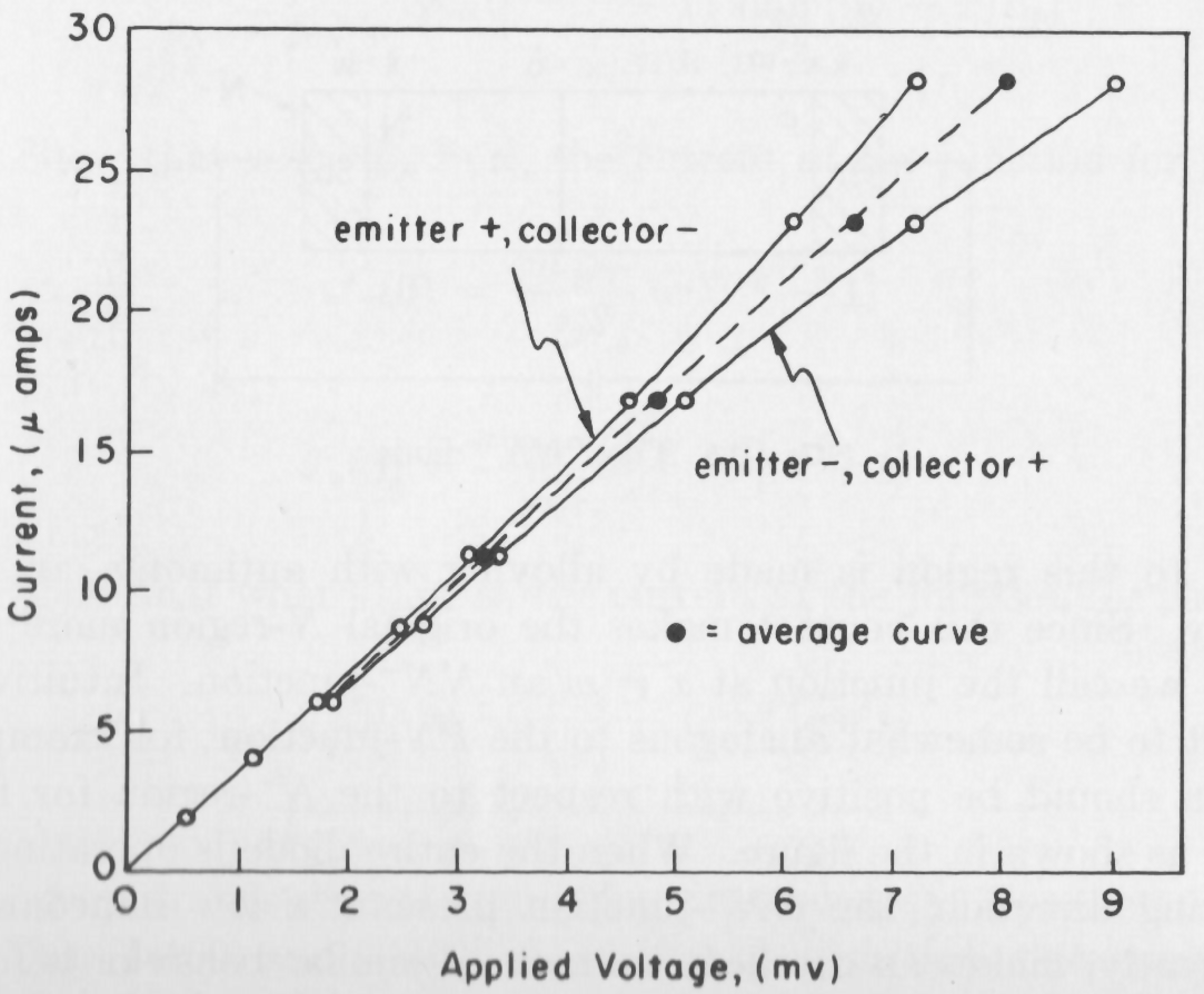


FIG. 4-15. Experimental data used in the determination of the base width. (From D. Long, Journ. Appl. Phys. **28**, 1219, 1957, with permission.)

If a voltage $\mathcal{V}$ is applied between the emitter and collector of a transistor, Eq. (4-93) is satisfied and I should vary linearly with $\mathcal{V}$. Figure 4-15 shows the experimental results of Long (1957) for a power transistor with the measurements made with both possible polarities. The average is seen to be a straight line, and the non-linearities of the individual curves are due to the fact that the emitter and collector areas are slightly different. The value of p_{0N} is obtained by measuring the resistivity of the original material and using

$$n_{0N} = 1/\rho e \mu_n \tag{1-24}$$

and

$$n_{0N} p_{0N} = n_i^2 \tag{3-24}$$

Hence, w can be determined from the slope of the $I - \mathcal{V}$ curve in the figure.

The curves for both the common base and common emitter connections, Figs. 4-11 and 4-13, indicate that the low-level theory developed above neglects to account for the variation of transistor properties with current. Later in this chapter, approximate methods of treating high level injection will be considered.

4-4 The PNN⁻ Diode and Transistor

Another low-level type of diode which is of practical importance is illustrated in Fig. 4-16, which shows a thin N-region of width w. The contact to this region is made by alloying with antimony, an N-type impurity. Since this contact makes the original N-region more heavily N-type, we call the junction at $x = w$ an NN^--junction. Intuitively we expect it to be somewhat analogous to the PN-junction; for example, the N-region should be positive with respect to the N^--region for forward biasing, as shown in the figure. When the entire diode is operating in the conducting direction, the NN^--junction presents a low impedance and consequently, makes an excellent contact. A similar behavior is found in the analogous P^+P-junction.

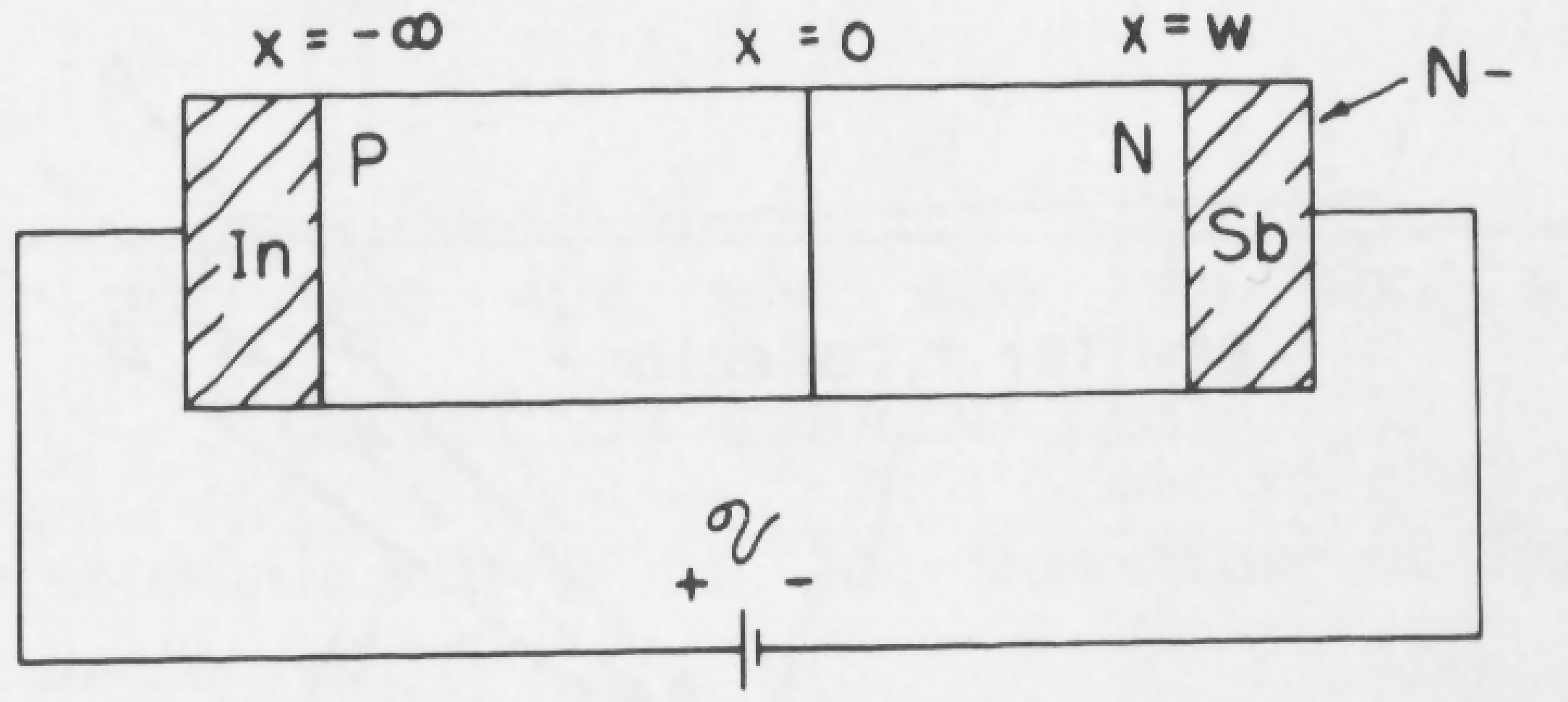


FIG. 4-16. The PNN^- diode.

The mathematical treatment of the PNN^- diode is very much like that of Section 4-2. The principal change is the boundary condition at $x = w$. By analogy with the PN-junction, the NN^--junction would be expected to be injecting a large electron current into the N^--region, and the hole current to be virtually zero.

Since

$$J_p = -e\,D_p \frac{dp}{dx} \tag{4-51}$$

this means that

$$\frac{dp}{dx} = 0 \tag{4-95}$$

and

$$-Ae^{-w/L_p} + Be^{w/L_p} = 0 \tag{4-96}$$

at $x = w$.

PROBLEM 4-21

(a) Combine Eq. (4-96) with Eq. (4-57) to show that for the N-region of the NN^--junction

$$p_N - p_{0N} = \frac{p_{0N}(e^{e\mathcal{V}/kT} - 1)\cosh\,[(w - x)/L_p]}{\cosh\,(w/L_p)} \tag{4-97}$$

(b) In the PN-junction of Fig. 4-7, let the width of the N region be w. Show that Eq. (4-59) is replaced by

$$p_N - p_{0N} = \frac{p_{0N}(e^{e\mathcal{V}/kT} - 1)\sinh\,[(w - x)L_p]}{\sinh\,(w/L_p)} \tag{4-98}$$

(c) Show that when $L_p \gg w$, the current at the junction for part (a) above is

$$J_p(0) = \frac{ew\,p_{0N}}{\tau_p}(e^{e\mathcal{V}/kT} - 1) \tag{4-99}$$

and for part (b) is

$$J_p(0) = \frac{e\,p_{0N}\sqrt{D_p}}{\sqrt{\tau_p}}(e^{e\mathcal{V}/kT} - 1) \tag{4-100}$$

(d) Show that when $L_p \ll w$, the current at the junction for part (a) is

$$J_p(0) = \frac{e\,p_{0N}\sqrt{D_p}}{\sqrt{\tau_p}}(e^{e\mathcal{V}/kT} - 1) \tag{4-101}$$

This problem has several interesting consequences. For example, comparing Eqs. (4-100) and (4-101) shows that the junction characteristics are independent of the contact at $x = w$ if the dimensions of the N-region are large compared to the diffusion length. The lifetime of carriers in a junction can be estimated by plotting a curve of I vs $\mathcal{V}$ and fitting it to either Eq. (4-99) or Eq. (4-100), using known values of w, p_{0N}, and D_p. Although this technique is not as accurate as the methods to be described, it is often the only one applicable to material on which junctions have been formed.

To consider now a transistor using an NN^- contact on the collector, we must extend our previous discussion to three dimensions. The equations of continuity [Eqs. (4-41) and (4-42)] can be extended to three dimensions without difficulty. Let us follow the method of Shockley (1950)[5] by considering a long rectangular rod of N-type material with dimensions $L \times 2B \times 2C$, as shown in Fig. 4-17. By

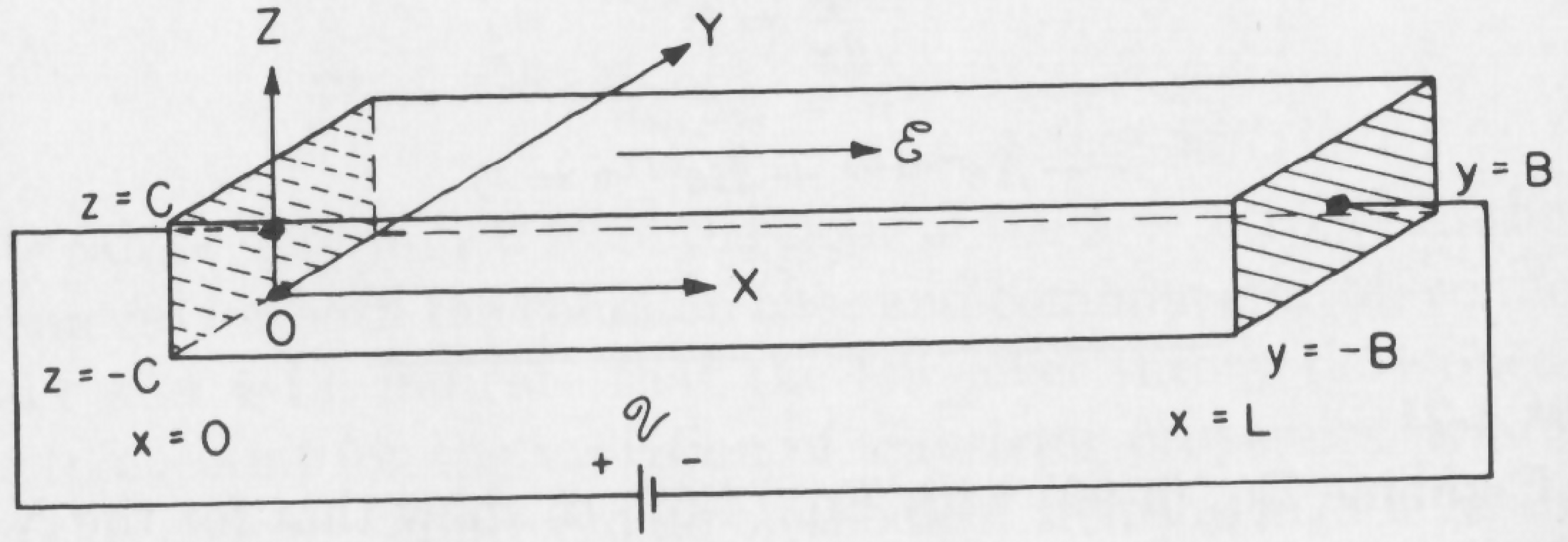


FIG. 4-17. A semiconducting filament. (From G.G.E. Low, *Proc. Phys. Soc.* **B68,** 310, 1955, with permission.)

$$\boldsymbol{\mathcal{E}} = -\textbf{grad}\ \mathcal{V} \tag{4-7}$$

Ohm's law is written $\quad \mathbf{J} = \sigma\boldsymbol{\mathcal{E}} = -\sigma\ \textbf{grad}\ \mathcal{V} \qquad$ (4-102)

The equation which corresponds to

$$J = (\sigma/e)\frac{d\bar{\mu}}{dx} \tag{2-102}$$

is thus $\quad \mathbf{J} = (\sigma/e)\ \textbf{grad}\ \bar{\mu}$

The continuity equation for holes [Eq. (4-41)] then becomes, in three-dimensional form,

$$\frac{\partial p}{\partial t} = \frac{-(p-p_0)}{\tau_p} - \frac{1}{e}\,\text{div}\ \mathbf{J}_p \tag{4-103}$$

(where we omit the subscript N we have previously used). In words, this equation states that the time rate of change of the hole concentration in a given volume is the sum of two contributions: the loss due to recombination and the net flow across the outer surface (using the interpretation of the divergence as developed in the appendix).

A voltage $\mathcal{V}$ is applied across the ends of the filament in Fig. 4-17, and if the contacts are carefully made, the resulting field $\mathcal{E}$ will be parallel to the x-axis and uniform. The hole current density $\boldsymbol{J}_p$ can then be expressed in component form as

$$\left.\begin{aligned} J_{px} &= pe\,\mu_p\mathcal{E} - e\,D_p\frac{\partial p}{\partial x} \\ J_{py} &= -e\,D_p\frac{\partial p}{\partial y} \\ J_{pz} &= -e\,D_p\frac{\partial p}{\partial z} \end{aligned}\right\} \tag{4-104}$$

Substituting these into the three-dimensional continuity equation [Eq. (4-103)] gives

$$\frac{\partial}{\partial t}(p-p_0) = \frac{-(p-p_0)}{\tau_p} - \mu_p\mathcal{E}\frac{\partial}{\partial x}(p-p_0) + D_p\nabla^2(p-p_0) \tag{4-105}$$

This equation has a solution

$$(p-p_0) = Ae^{-\nu t - ax}\cos by\cos cz \tag{4-106}$$

where A is an integration constant and a, b, c, and ν are constants related by

$$\nu + D_p a^2 + \mu_p\mathcal{E}a = \frac{1}{\tau_p} + D_p(b^2+c^2) \tag{4-107}$$

At the external surfaces we have a sudden discontinuity in the crystal lattice, and regions of this type are known to produce rapid recombination. We shall assume, as before, that the recombination rate is proportional to the excess hole concentration $(p - p_0)$; hence, the *recombination current density* J_r at the surface should be of the form

$$J_r = e\,s(p-p_0) \tag{4-108}$$

where s is a proportionality constant called the *surface recombination velocity*. Since J_r at $y = B$ can also be expressed

$$J_r = -e\,D_p\frac{\partial}{\partial y}(p-p_0)$$

we have

$$-e\,D_p\frac{\partial p}{\partial y} = (p-p_0)e\,s \tag{4-109}$$

At $y = -B$, J_r will have the opposite sign, and the boundary conditions on the four surfaces are

$$\left.\begin{aligned} D_p\frac{\partial p}{\partial y} &= \pm s(p - p_0), \qquad y = \pm B \\ D_p\frac{\partial p}{\partial z} &= \pm s(p - p_0), \qquad z = \pm C \end{aligned}\right\} \tag{4-110}$$

PROBLEM 4-22 Substituting Eq. (4-106) into Eq. (4-110), show that

$$\left.\begin{aligned} sB/D_p &= \eta \tan \eta \\ sC/D_p &= \zeta \tan \zeta \end{aligned}\right\} \tag{4-111}$$

where $\eta = bB, \qquad \zeta = cC$

These two transcendental equations must be solved by approximate methods, and each has an infinite number of roots.

PROBLEM 4-23 For a sandblasted germanium surface, s is about 10^4 cm/sec. Find the root s of lowest numerical value of Eq. (4-111) for a filament of 1 mm cross section.

Let us consider now the situation in which the hole density is independent of x, so that $a = 0$. Then Eq. (4-107) becomes

$$\nu = \frac{1}{\tau_f} = \frac{1}{\tau_p} + D_p\left(\frac{\eta^2}{B^2} + \frac{\zeta^2}{C^2}\right) \tag{4-112}$$

where τ_f is called the *filament lifetime.* The coefficient of D_p involves B, C, and s, which are surface properties, and hence this term can be regarded as the *surface lifetime* $1/\tau_s$, so

$$\frac{1}{\tau_f} = \frac{1}{\tau_p} + \frac{1}{\tau_s} \tag{4-113}$$

PROBLEM 4-24

(a) Show that when $s \rightarrow \infty$ and $p = p_0$

$$\frac{1}{\tau_s} = \frac{\pi^2 D_p}{4}\left(\frac{1}{B^2} + \frac{1}{C^2}\right)$$

(b) Show that when $s \rightarrow 0$

$$\frac{1}{\tau_s} = s\left(\frac{1}{B} + \frac{1}{C}\right)$$

PROBLEM 4-25

(a) Consider the continuity equation with injection existing but with $J_p = 0$. Solve this equation and show how the constant in the exponent of your solution can be interpreted in terms of a probability.

(b) How does this interpretation correspond with Eq. (4-113)?

If we now remove the restriction on a, but instead establish steady-state conditions, then $\nu = 0$ and, from Eq. (4-107),

$$a = \frac{-\mu_p\mathcal{E} \pm (\mu_p^2\mathcal{E}^2 + 4D_p/\tau_f)^{1/2}}{2D_p}$$

If the diffusion is small compared to the effect of the field, then $\sqrt{D_p/\tau_f} \ll \mu_p\mathcal{E}$, and a binomial expansion gives

$$a \sim 1/\mu_p\mathcal{E}\,\tau_p$$

The quantity $1/a$, which has the dimensions of length, is called by Low (1955)[10] the *downstream diffusion length*, since the field assists the motion of holes. If the injection efficiency γ is low and the field is strong, it is possible to have an effect called *exclusion*, which is the opposite of injection. Exclusion means that holes are moved away from the contact at $x = 0$ so fast that neither the injected nor the thermally excited holes can completely replace them. The result is to have a deficiency of holes at this contact, rather than an excess as for injection, and $(p - p_0)$ is a negative number. If we evaluate A of Eq. (4-106) at $x = 0$ we have

$$A = (p - p_0)_0/\cos by \cos cz$$

and $\quad (p - p_0) = (p - p_0)|_{x=0}e^{-ax}$

Hence, p will decay (injection) or rise (exclusion) exponentially to p_0, as shown in Fig. 4-18, depending on whether $(p - p_0)|_{x=0}$ is positive or negative. It should be observed that injection is associated with a high γ and exclusion with a low γ. If we now consider high and low γ contacts at $x = L$, we have analogous effects. For example, *accumulation* can occur when γ is low because the holes cannot be removed fast enough. On the other hand, *extraction* will occur for high γ because the holes are removed faster than they can be replaced.

A PNN^- transistor using this ac-

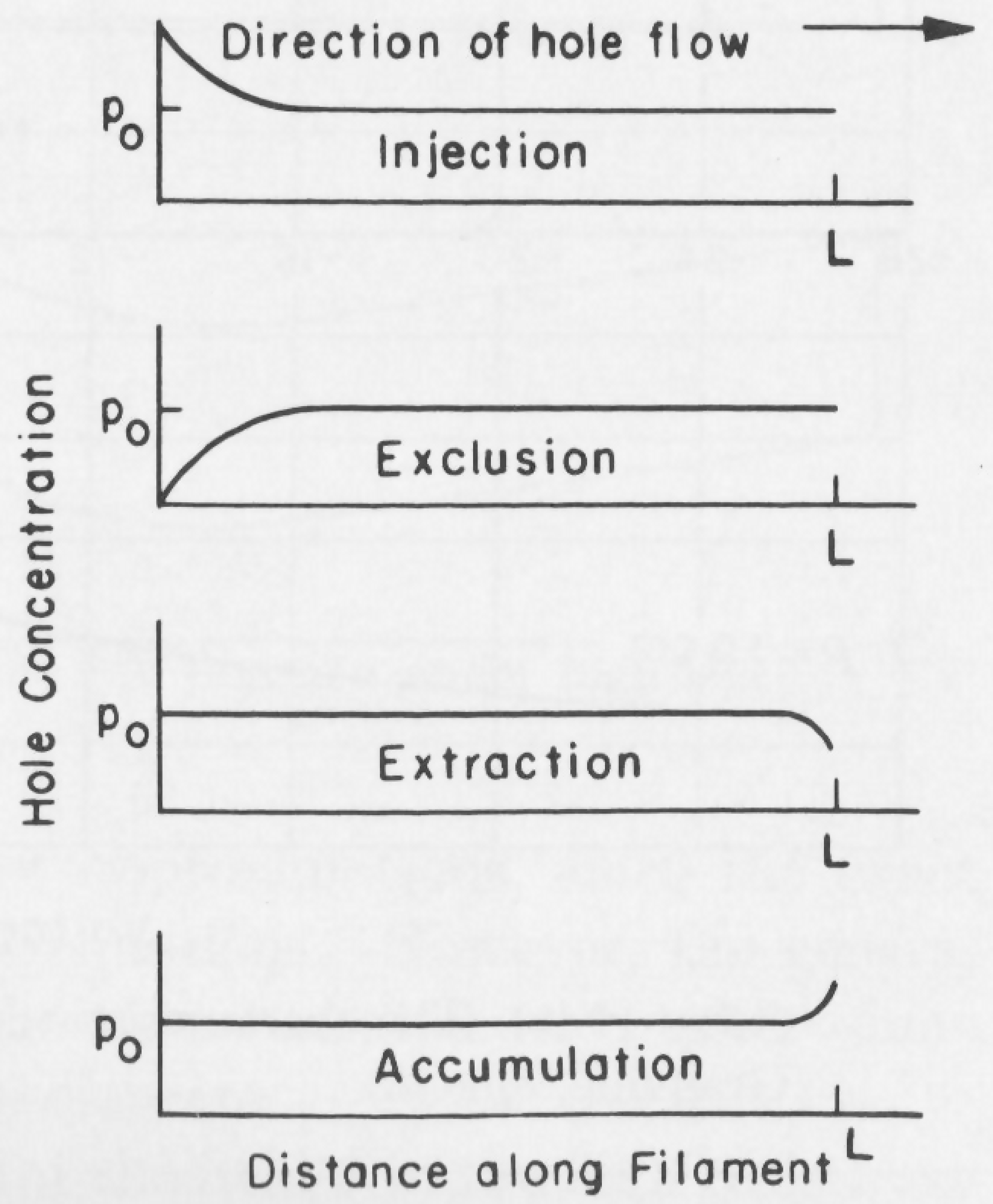


FIG. 4-18. The four types of carrier transfer processes in a semiconductor. (From Low, *op. cit.*)

cumulation property has been reported by Granville (1956)[11]. Figure 4-19 shows a cross section and a plot of the current vs voltage of the NN^--junction for a 3-ohm-cm germanium base. Note that the behavior near the origin is almost ohmic, and we would expect the reasoning of Problem 4-15(c) to apply. The action at the NN^--junction is based on the fact that N-region has a relatively high resistivity, and hence a comparatively large field exists in it. Since γ is low, holes pile up in this region and accumulation occurs. The accumulation, in turn, brings a large electron current to flow into the N-region, and the junction is essentially ohmic.

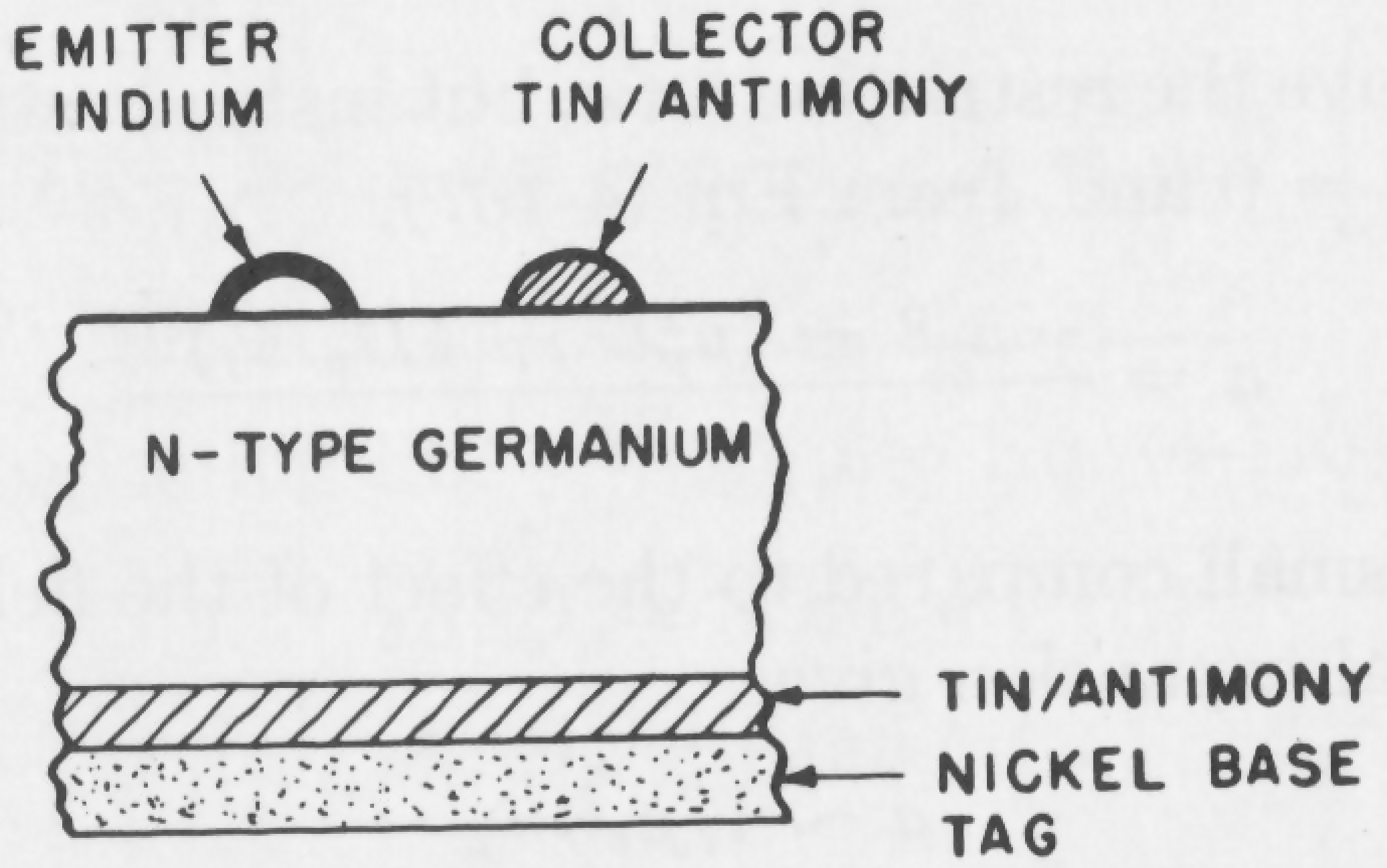


FIG. 4-19 (a). A PNN^- transistor. (From J. W. Granville, J. Elec. 1, 565, 1956, with permission.)

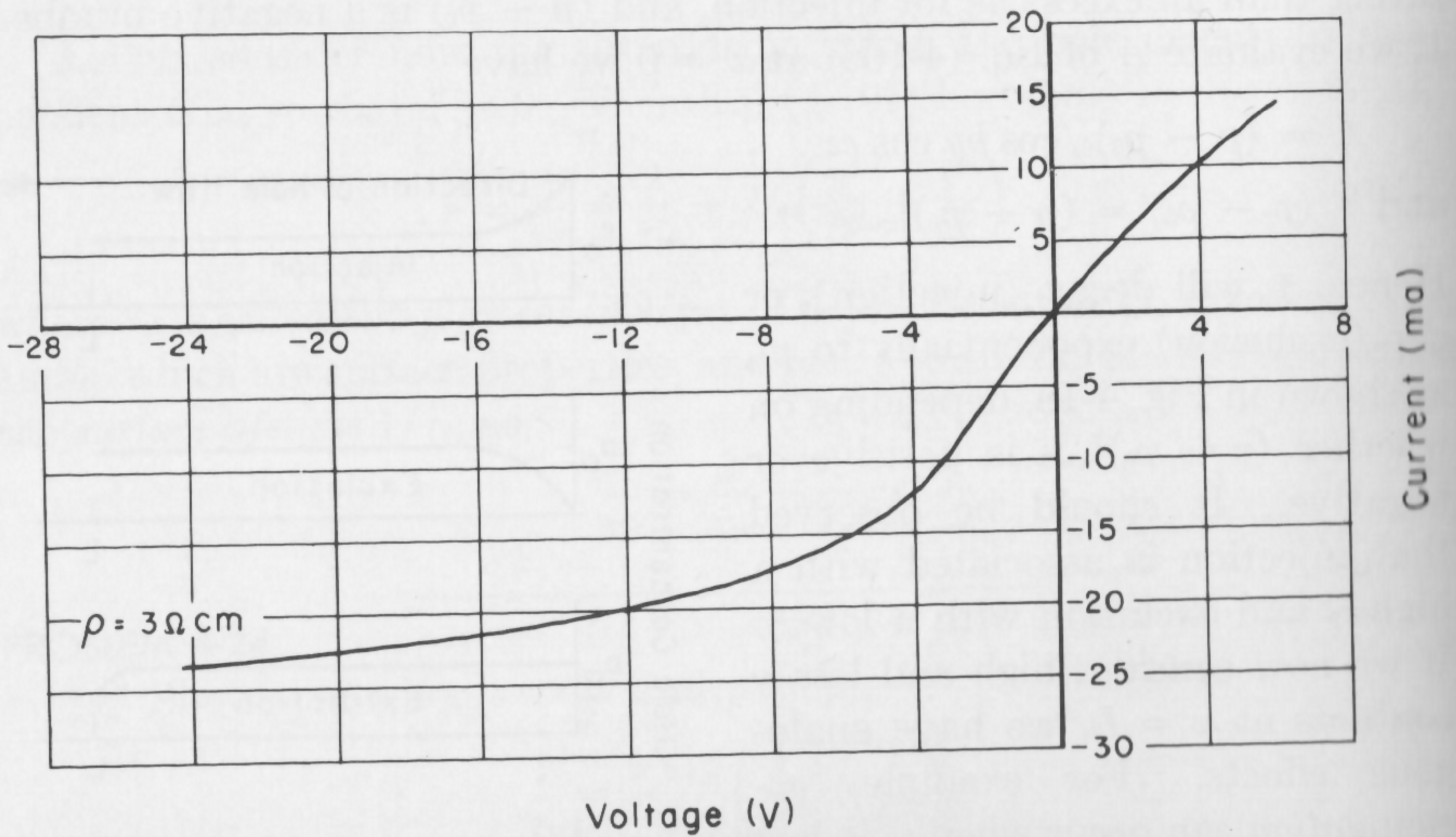


FIG. 4-19 (b). The characteristic curve for the PNN^- transistor. (From Granville, *op. cit.*)

4-5 Junction Devices Under High-Level Injection

The diode at high levels of injection is considerably more complicated than the low-level diode, because the applied voltage will produce an internal field $\mathcal{E} = -d\mathcal{V}/dx$ which we can no longer neglect in the higher resistivity region of the junction; that is, in the N-region. The P-region is so heavily doped that its resistivity is extremely low and the bulk of the applied voltage appears across the junction, with some part of it being across the N-region. Since $d\mathcal{V}/dx \neq 0$ in this region, the energy level diagram, shown in Fig. 4-20, is also more involved than in the low-level case. The treatment which follows incorporates a number of approximations, since the exact analysis involves non-linear differential equations. However, the general features of high-level operation can be demonstrated, and after some simple calculations, we shall discuss the diagram.

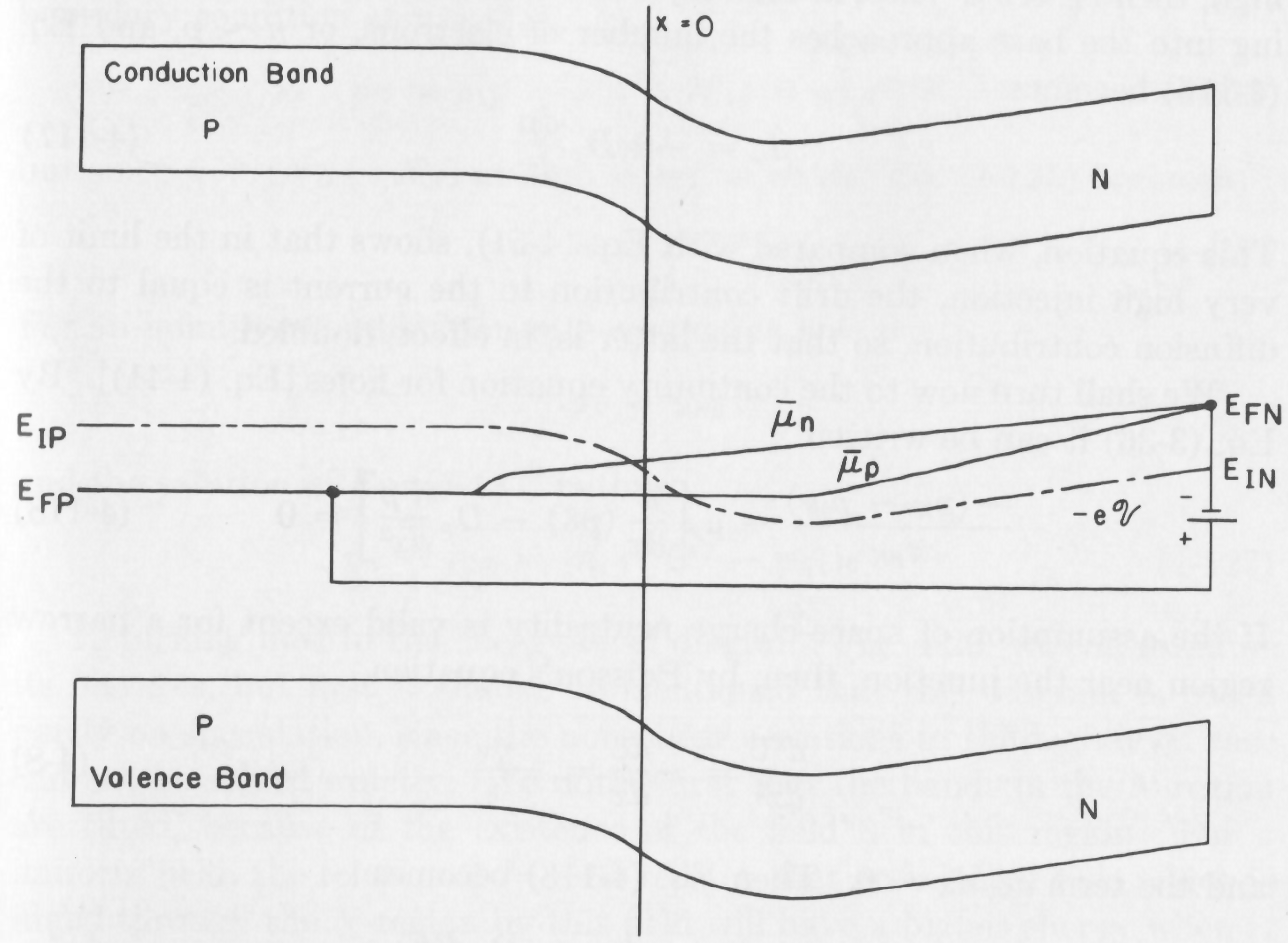


FIG. 4-20. Energy level diagram for the PN junction under high level injection.

Let us solve Eqs. (3-36) and (3-40) for the field $\mathcal{E}$ in the high resistivity N-region, and equate the results.

$$\mathcal{E} = \frac{J_p + e\,D_p\,dp/dx}{p\,\mu_p} = \frac{J_n - e\,D_n\,dn/dx}{n\,\mu_n} \qquad (4\text{-}114)$$

Since the donors and acceptors are fully ionized, by Eq. (3-7),

$$n + N_A = p + N_D$$

and

$$dn/dx = dp/dx \tag{4-115}$$

Putting Eq. (4-115) into Eq. (4-114) yields

$$\frac{dp}{dx} = \frac{(J_n p - J_p n\, c)}{ec\, D_p\, (n + p)} \tag{4-116}$$

where $c = \mu_n/\mu_p$ is called the *mobility ratio.* If the injection efficiency γ is high, then $J_n \ll J_p$. Also, at high-injection levels, the number of holes coming into the base approaches the number of electrons, or $n \sim p$, and Eq. (4-116) becomes

$$J_p = -2eD_p \frac{dp}{dx} \tag{4-117}$$

This equation, when compared with Eq. (4-51), shows that in the limit of very high injection, the drift contribution to the current is equal to the diffusion contribution, so that the latter is, in effect, doubled.

We shall turn now to the continuity equation for holes [Eq. (4-41)]. By Eq. (3-36) it can be written

$$\frac{-(p_N - p_{0N})}{\tau_{pN}} - \mu_p\left[\frac{d}{dx}(p\mathcal{E}) - D_p \frac{d^2p}{dx^2}\right] = 0 \tag{4-118}$$

If the assumption of space-charge neutrality is valid except for a narrow region near the junction, then, by Poisson's equation,

$$\frac{d^2\mathcal{V}}{dx^2} = -\frac{d\mathcal{E}}{dx} = -\frac{qv}{\epsilon} \tag{4-8}$$

and the term $d\mathcal{E}/dx = 0$. Then Eq. (4-118) becomes

$$\frac{-(p_N - p_{0N})}{\mu_p \tau_{pN}} - \mathcal{E}\frac{dp_N}{dx} + \frac{D_p}{\mu_p}\frac{d^2p_N}{dx^2} = 0 \tag{4-119}$$

Similarly,

$$\frac{-(n_N - n_{0N})}{\mu_n \tau_{nN}} + \mathcal{E}\frac{dn_N}{dx} + \frac{D_n}{\mu_n}\frac{d^2n_N}{dx^2} = 0 \tag{4-120}$$

For space-charge neutrality

$$p_N - p_{0N} = n_N - n_{0N} \tag{4-121}$$

and we will make the further assumption that the lifetimes are equal

$$\tau_{pN} = \tau_{nN} \tag{4-122}$$

(This assumption will be discussed when we consider trapping.) Adding Eqs. (4-119) and (4-120) and using Eqs. (4-121) and (4-122), we obtain

$$\frac{-(p - p_{0N})}{\tau_{pN}}\left(\frac{1}{\mu_p} + \frac{1}{\mu_n}\right) + \frac{2D_p}{\mu_p}\frac{d^2p_N}{dx^2} = 0$$

or

$$\frac{d^2p_N}{dx^2} - \frac{(p - p_{0N})}{L_p'^2} = 0 \tag{4-123}$$

where

$$L_p' = \frac{2c}{c + 1} L_p \tag{4-124}$$

is called the *high-level diffusion length.* Note that $L_p' = L_p$ for $\mu_n = \mu_p$. The boundary condition at $x = 0$ is

$$pn = p(p + N_D - N_A) = n_i^2\, e^{e\mathcal{V}/kT} \tag{4-125}$$

but since $p \gg (N_D - N_A)$ at high injection levels, Eq. (4-125) becomes

$$p = n_i\, e^{e\mathcal{V}/2kT} \tag{4-126}$$

For an infinite recombination rate contact at $x = \infty$,

$$p_N = p_{N0}$$

and the solution of Eq. (4-123) is thus

$$p_N - p_{N0} = (n_i\, e^{e\mathcal{V}/2kT} - p_{N0})e^{-x/L'_p} \tag{4-127}$$

Returning now to the energy-level diagram, Fig. 4-20, we can examine its features, but first it should be mentioned that this diagram is based partly on speculation, since the non-linear equations in the high-level case can not be solved exactly. We notice first that the bands in the N-region are tilted, because of the existence of the field $\mathcal{E}$ in this region. For a uniform field, the relation $\mathcal{V} = -\mathcal{E}x$ tells us that an injected hole which is urged through the N-region by this field will have a higher energy when it reaches the right-hand contact, since work must be done on it to move it the other way. Hence, this end of the N-region is at a higher energy and the bands bend upward. An intuitive way of looking at this is to think of electrons as marbles and holes as bubbles. Unconstrained holes will move from left to right with the N-region tilted as shown and electrons will tend to go in the other direction. With this simple analogy, the space-charge region has the opposite effect: it represents a barrier to both kinds of carrier, as we already know it to be. Another interesting point shown in the figure is the separation of $\bar{\mu}_n$ and $\bar{\mu}_p$ right up to the electrical contact. Because of the high level of injection, both the electron and hole concentrations are different from their equilibrium value throughout the entire

N-region. We note that the implicit assumption we made in Eq. (4-125), namely, that the applied potential $-e\mathcal{V}$ represents $(\bar{\mu}_n - \bar{\mu}_p)$ measured at $x = 0$, is only approximately correct.

If we place the N-region contact at $x = w$, then by using the same calculations as were used in Problem 4-21, we obtain

$$p_N - p_{0N} = \frac{\sinh\,[(w - x)/L_p'](n_i\, e^{e\mathcal{V}/2kT} - p_{0N})}{\sinh w/L_p'}$$

Using Eq. (4-51) and the condition that $w \ll L_p'$ gives

$$J_p \sim J = (2e\, D_p\, n_i/w)e^{e\mathcal{V}/2kT} \tag{4-128}$$

This relation was given by Hall (1952)[12], and has been experimentally verified by Long (1956)[13], using germanium diodes. The samples for these experiments were made by fusing pure indium or a 99 per cent In-1 per cent Ga alloy into one side of a thin wafer of N-type Ge and pure tin into the opposite side. In Fig. 4-21 is shown the current vs voltage curve for a diode made from 35 ohm-cm resistivity material. It is seen that at high current densities the slope of the curve gradually approaches $e\mathcal{V}/2kT$, as we would expect from Eq. (4-128). The agreement was not as close for diodes made from lower resistivity material, and it is believed that this may be due in part to the voltage drop across the tin contact at high currents.

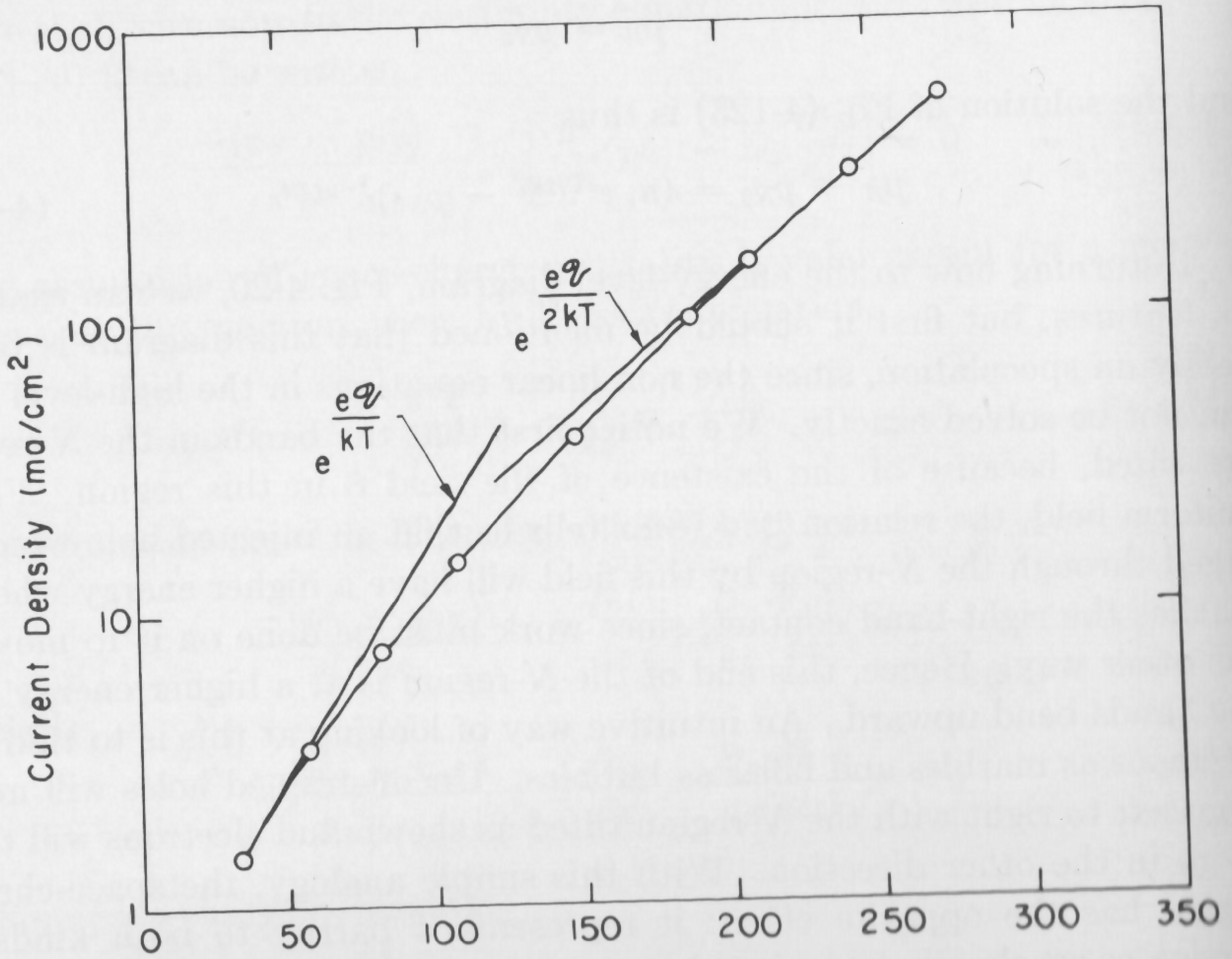


FIG. 4-21. The transition from low to high level injection in a diode. (From D. Long, Honeywell Research Center unpublished report, 1956, with permission.)

The electron current injected into the P-region of the diode is given approximately by Eq. (4-63).

$$J_n \sim (e\, D_n\, n_{0P}/L_n)e^{e\mathcal{V}/kT}$$

The fraction of the current which is composed of *electrons* is then:

$$J_n/J = \frac{D_n\, n_{0P}\, w}{2D_p\, n_i\, L_n}\, e^{e\mathcal{V}/2kT}$$

$$= \frac{D_n\, n_{0P}\, w^2}{4D_p^2\, n_i^2\, L_n}\, J \tag{4-129}$$

This quantity increases with J, so the fraction of the current composed of holes — that is, the injection efficiency γ — decreases with current.

The theory of transistors undergoing high-level injection has been given by Webster (1954)[14], Rittner (1954)[15], Misawa (1956)[16], and Matz (1960)[17]. Following Rittner, Eq. (4-114) is written in the form

$$dx = \frac{-e\, D_p}{J_p}\left\{\frac{(2p + N)\, dp}{p[1 - (J_n/cJ_p)] + N}\right\} \tag{4-130}$$

where we have used

$$n - p = N_D - N_A = N$$

and

$$dp/dx = dn/dx \tag{4-115}$$

If the injection efficiency is reasonably high, most of the current across the emitter junction is carried by holes, and $J_n/cJ_p \ll 1$. Hence, Eq. (4-130) reduces to

$$dx = \frac{-e\, D_p}{J_p}\left[\frac{(2p + N)\, dp}{p + N}\right]$$

Upon integration, this gives

$$x + C = \frac{-e\, D_p\, N}{J_p}\left[\ln\,(p + N) - \frac{2}{N}\,(p + N)\right] \tag{4-131}$$

We evaluate C by the boundary condition obtained from Eq. (4-57).

$$p = p_{0N}\, e^{e\mathcal{V}_E/kT} = p_E \qquad \text{at } x = 0 \tag{4-132}$$

where p_E is the hole density at the emitter junction, and

$$x = \frac{q\, D_p\, N}{J_p}\left\{\frac{2(p_E - p)}{N} - \ln\left[\frac{(1 + p_E/N)}{(1 + p/N)}\right]\right\} \tag{4-133}$$

By putting the boundary condition

$$p = p_{0N}\, e^{-e\mathcal{V}_c/kT} = p_C \qquad \text{at } x = w \tag{4-69}$$

into Eq. (4-133), we get the hole current in the base as

$$J_{pE} = J_{pC} = J_p = \frac{q\, D_p\, N}{w}\left\{\frac{2(p_E - p_C)}{N} - \ln\left[\frac{1 + p_E/N}{1 + p_C/N}\right]\right\} \tag{4-134}$$

For high-level injection, the boundary condition at $x = 0$

$$n_P = n_{0P}\, e^{e\mathcal{V}_E/kT}$$

has been shown by Misawa (1955)[16] to be

$$n_P = n_{0P}\, e^{\mathcal{V}_E/kT}\,(1 + p_{0N}\, e^{e\mathcal{V}_E/kT}/n_{0N}) \tag{4-135}$$

To establish Eq. (4-135), we assume space-charge neutrality in the base

$$\Delta n = \Delta p$$

or

$$n_N - n_{0N} = p_N - p_{0N} \tag{4-136}$$

In Fig. 4-6(a), the applied voltage $\mathcal{V}$ corresponds to the difference in electrochemical potentials at the junction, and from Eq. (4-55)

$$p_N n_N = p_{0N} n_{0N}\, e^{e\mathcal{V}_E/kT} \tag{4-137}$$

for the low-level situation. For the case of high-level injection, as shown in Fig. 4-20, this correspondence is no longer true, but can be considered to be a reasonable approximation. Combining Eqs. (4-136) and (4-137) gives

$$(p_{0N}/n_{0N})\, e^{e\mathcal{V}_E/kT} = (p_N/n_{0N}{}^2)(n_{0N} + p_N - p_{0N}) \tag{4-138}$$

If we neglect p_{0N} in comparison to n_{0N}, Eq. (4-138) becomes

$$(p_{0N}/n_{0N})\, e^{e\mathcal{V}_E/kT} = (p_N/n_{0N})^2 + (p_N/n_{0N}) \tag{4-139}$$

Since this equation is being applied at the emitter junction, $p_N = p_E$. For convenience, let

$$P = p_E/n_{0N}, \qquad V = (p_{0N}/n_{0N})e^{e\mathcal{V}_E/kT}$$

Then Eq. (4-139) becomes

$$V = P^2 + P = P(1 + P) \tag{4-140}$$

whereas Eq. (4-132) is

$$V = P \tag{4-132}$$

and the two equations do not agree. Let us assume that Eq. (4-140) is the correct high-level boundary condition for the moment.

Next, let $\mathcal{V}_c$ be large enough so that $e^{-e\mathcal{V}_C/kT}$ is negligible, and using $p_{0N} \ll n_{0N}$ again, Eq. (4-134) becomes

$$J_{pE} = \frac{q\, D_p\, N}{w}\,[2P - \ln(1 + P)] \tag{4-141}$$

The emitter electron current, from Eq. (4-75), is

$$J_{nE} = \frac{e\, D_n\, n_{0P}}{L_n}\,[(n_{0N}/p_{0N})V - 1] \tag{4-142}$$

PROBLEM 4-26

(a) Show that

$$\frac{\partial J_{nE}}{\partial J_{pE}} = \frac{n_{0P}\, D_n\, w}{p_{0N}\, D_p\, L_n}\,(1 + P) \tag{4-143}$$

using Eqs. (4-140), (4-141), and (4-142).

(b) Show that Eq. (4-143) is also obtained from Eqs. (4-135), (4-132), and (4-141).

This problem shows that the two sets of boundary conditions

$$P = V, \qquad n_E = n_{0P}\, e^{e\mathcal{V}_E/kT}\,(1 + P_E/N)$$

and

$$P^2 + P = V, \qquad n_E = n_{0P}\, e^{e\mathcal{V}_E/kT}$$

are equivalent. In what follows, we shall use Eqs. (4-135) and (4-140).

The current J_r lost in the base through recombination has a surface and a volume component, and can be expressed approximately by

$$J_r = es(p_N - p_{0N}) + ew(p_N - p_{0N})/\tau_p \tag{4-144}$$

PROBLEM 4-27 Justify Eq. (4-144).

The emitter hole current is then diminished by an amount J_r by the time it reaches the collector, so that

$$J_{pE} = J_{pc} + J_r \tag{4-145}$$

PROBLEM 4-28

(a) Show that the electron currents in the high-level transistor are

$$J_{nE} = \frac{e\, D_n}{L_n}\,[n_E(1 + p_E/N) - n_{0P}] \tag{4-146}$$

and

$$J_{nC} = \frac{-e\, D_n}{L_n}\,(n_C - n_{0P}) \tag{4-147}$$

where

$$n_C = n_{0P}\, e^{-e\mathcal{V}_c/kT}$$

(b) Using Eqs. (4-80), (4-134), (4-145), (4-146), (4-147), and the assumption that $\mathcal{V}_c$ is large enough so that $e^{-e\mathcal{V}_C/kT} \sim 0$, show that

$$1 - \alpha = \frac{sp_{0N} + (wp_{0N}/\tau_p) + (D_n n_{0P}/L_n)[1 + (2p_{0N}/N)e^{e\mathcal{V}_E/kT}]}{\left[\begin{array}{l}(2D_p p_{0N}/w) - \{D_p p_{0N}/w[1 + (p_{0N}/N)e^{e\mathcal{V}_E/kT}]\} \\ \qquad + (D_n n_{0P}/L_n)[1 + (2p_{0N}/N)e^{e\mathcal{V}_E/kT}]\end{array}\right]} \tag{4-148}$$

(c) Show that Eq. (4-148) reduces to

$$1 - \alpha = \frac{sw}{D_p} + \frac{w^2}{L_p^2} + \frac{D_n n_{0P} w}{D_p p_{0N} L_n} \tag{4-149}$$

when $\mathcal{V}_E$ is small (low injection) and when $n_{0P} \ll p_{0N}$ (well-doped emitter). Identify the terms on the right of Eq. (4-149) with surface and volume recombination and injection efficiency.

From Eq. (4-88), we see that for a transistor with $\alpha \sim 1.0$, the common emitter amplification is approximately

$$\alpha_{CE} = \frac{1}{1 - \alpha}$$

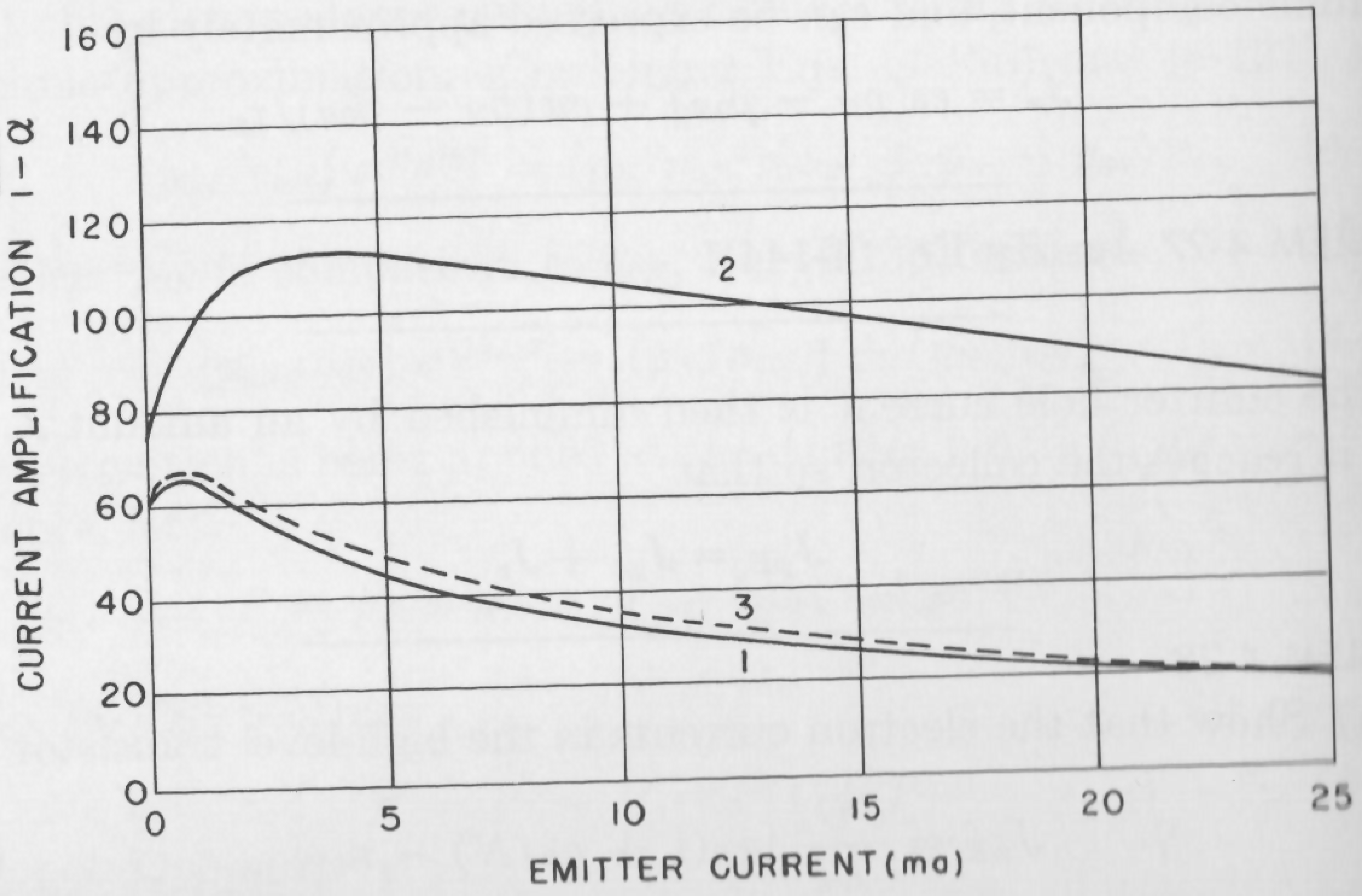


FIG. 4-22. Variation of amplification with emitter current for *PNP* transistors; see text for explanation. (From E. S. Rittner, Phys. Rev. **94,** 1161, 1954 and W. M. Webster, Proc. IRE **42,** 914, 1954, with permission.)

By plotting $1/(1 - \alpha)$ as a function of J_E with $\mathcal{V}_E$ as a parameter in Eqs. (4-134), (4-146), and (4-148), Rittner (1954)[15] obtains the two solid curves shown in Fig. 4-22. Curve 1 is for a transistor with about 10^{15} donors/cm^3 in the base and 10^{18} acceptors/cm^3 in the emitter, like the junction of Fig. 4-3, and curve 2 has the emitter doped to 10^{19} acceptors/cm^3. Curve 3 (dashed) shows the value of α_{CE} for an experimental transistor measured by Webster (1954). The resistivities and other physical constants of the transistor were chosen to be approximately the same for curves 1 and 3; consequently, they match at the starting point and also for their subsequent shapes.

4-6 Generation and Recombination Processes in Junction Devices

In the previous section, high-level devices were discussed under the assumption that the minority carrier lifetime is constant. We now wish to show that this lifetime depends in fact upon the injected carrier density, and that its variation can account for the change in the gain of some transistors at higher currents.

Let us consider first the theory of Hall (1952)[18] and of Shockley and Read (1952)[19] regarding the generation and recombination of carriers in junctions. In originally discussing lifetime and the concept of recombination, no mention was made of the mechanism through which this process can occur. The most obvious way would be a collision between an electron and a hole — this is called *direct recombination* — but we can show by an analogy that this is unlikely to occur. If two people are throwing basketballs with the intent of having one person's ball collide with the other, the probability of success is low. On the other hand, if they both are shooting for the same basket, at random, many collisions are bound to occur. In a semiconductor, a similar situation occurs. However, both electrons and holes can be caught by *recombination centers* or *traps*, which are defects such as missing and extra atoms or structural faults in the lattice. The energies associated with these defects are found experimentally to lie generally closer to the center of the forbidden gap than the impurity levels. For example, Burton et al. (1953)[20] have shown that copper, which is an *interstitial* impurity in germanium, has a trap level about 0.25 eV above the valence band. In Fig. 4-23 the basic processes associated with traps are shown. The four processes shown, and the names given them, are

(1) Electron falls into trap	electron capture
(2) Electron leaves trap	electron emission
(3) Electron falls into valence band	hole capture
(4) Electron leaves valence band	hole emission

Recombination then consists of process (1) followed by process (3); generation may be accomplished by process (4) followed by process (2).

Now we consider the effect of doping on the recombination process.

PROBLEM 4-29
(a) Consider a highly doped P-type sample with copper as the trap. From the position of the Fermi level, show that the traps are almost empty.

(b) Show that the traps become more and more occupied as the degree of doping shifts from heavy to light and then to intrinsic.

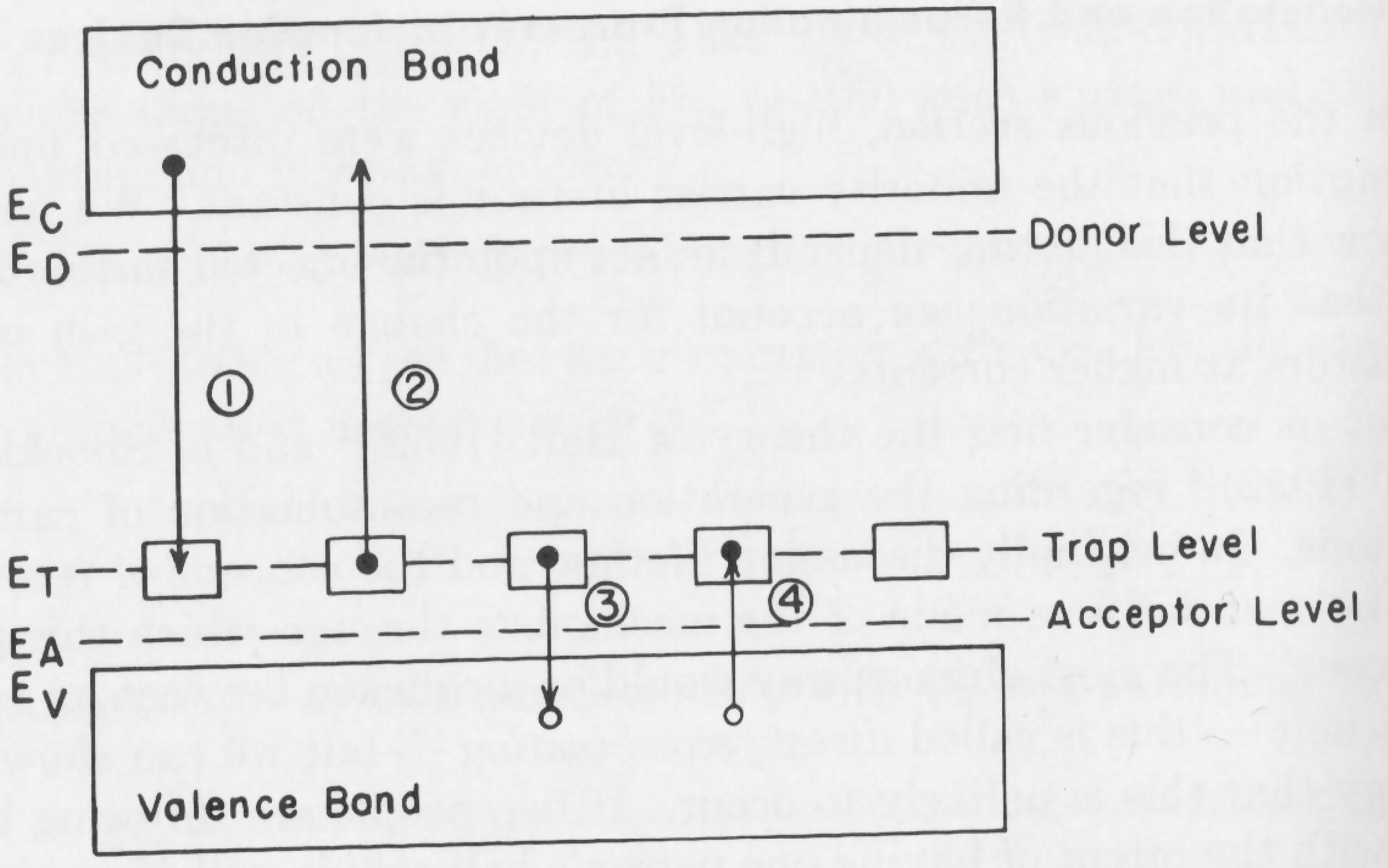


FIG. 4-23. Carrier emission and trapping; see text for explanation.

For the highly-doped sample with low trap occupancy, the recombination should occur quickly because of the large number of holes available. The lifetime of electrons in highly doped P-type material is defined as τ_{n0}, and represents a lower limit value. As the concentration decreases and the trap occupancy increases, the lifetime goes up also, reaching its maximum for intrinsic material. It then goes down for N-type material, and the lifetime of holes in highly doped N-type material is denoted by τ_{p0}.

Considering direct recombination for a moment, it would be expected that the recombination rate is proportional to both the electron and hole density; that is, it is a quadratic function of concentration. For the present situation, we assume a linear rate and write the lifetime as

$$\tau = \alpha n_0 + \beta p_0 + \gamma$$

where n_0 and p_0 are equilibrium concentrations and α, β, and γ are constants.

The constant γ is necessary to incorporate the dependence of the trap energy level. We shall show shortly that

$$\alpha = \frac{\tau_{p0}}{n_0 + p_0}, \qquad \beta = \frac{\tau_{n0}}{n_0 + p_0} \tag{4-150}$$

$$\gamma = \frac{\tau_{p0}\, n_1 + \tau_{n0}\, p_1}{n_0 + p_0}$$

where

$$n_1 = n_i\, e^{(E_T - E_I)/kT} \tag{4-151}$$

$$p_1 = n_i\, e^{(E_I - E_T)/kT}$$

and n_1 and p_1 are the densities of free carriers when the Fermi level equals the trap level E_T. Substituting, we obtain

$$\tau = \frac{\tau_{p0}(n_0 + n_1) + \tau_{n0}(p_0 + p_1)}{n_0 + p_0} \tag{4-152}$$

PROBLEM 4-30 Consider a highly doped P-type sample and show from Eq. (4-152) that $\tau = \tau_{n0}$. This verifies that Eqs. (4-150) are consistent with the definition of τ_{n0} given above.

Equation (4-152) has been verified experimentally by Hall (1952)[18], who found $\tau_{p0} = 100$ μsec, $\tau_{n0} = 10$ μsec, and $\tau_{\max} = 1200$ μsec in germanium.

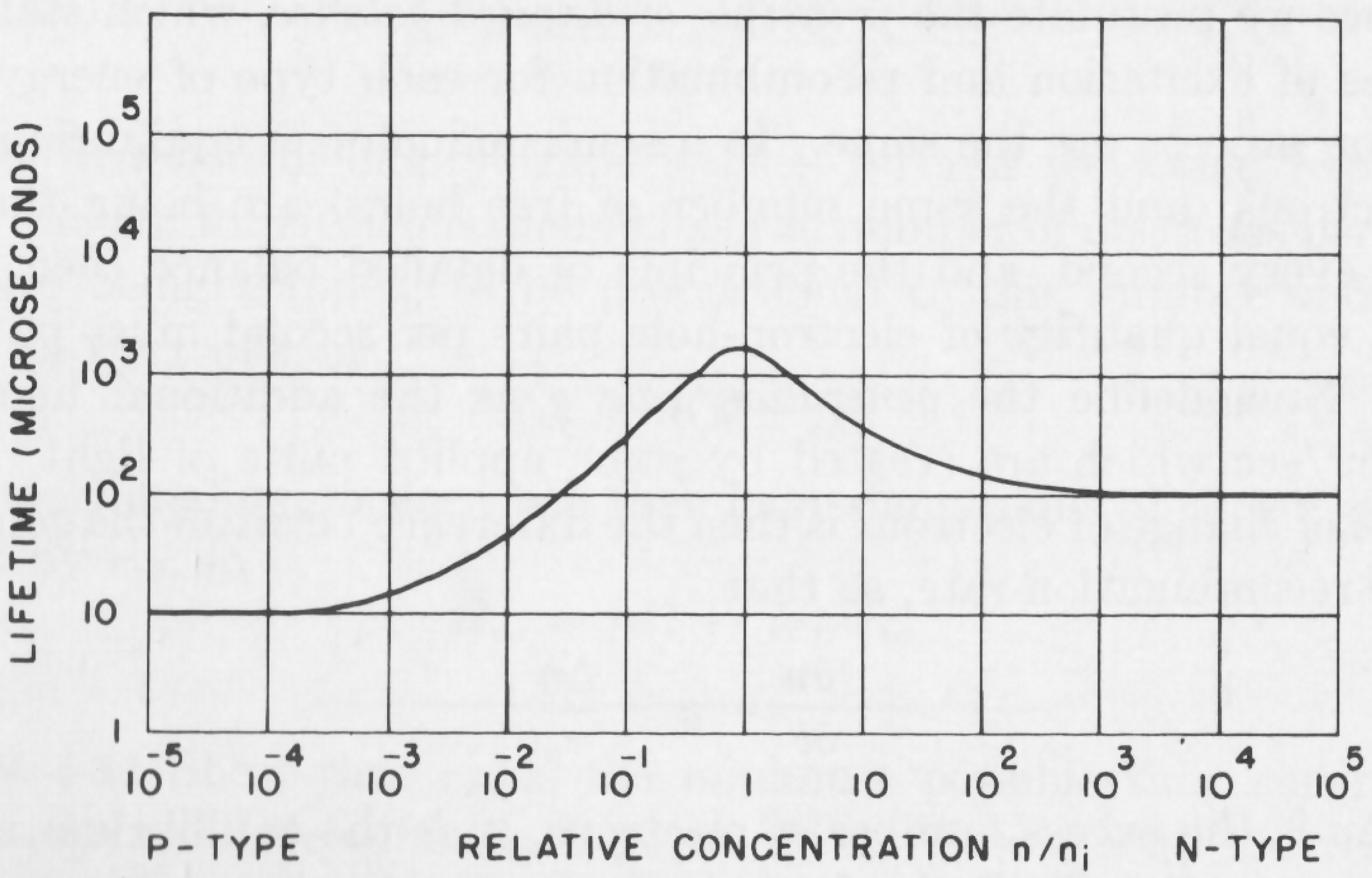


FIG. 4-24. Minority carrier lifetime as a function of concentration for germanium. (From R. N. Hall, Phys. Rev. **87**, 387, 1952, with permission.)

The curve of Fig. 4-24 shows his results for a number of germanium samples with different amount of N-type or P-type impurities.

PROBLEM 4-31

(a) Show that Eq. (4-152) can be put in the form

$$\tau = \frac{\tau_{p0}x + (\tau_{n0}/x) + c}{x + (1/x)} \tag{4-153}$$

where

$$x = n_0/n_i$$
$$c = a\tau_{p0} + \tau_{n0}/a$$
$$a = n_1/n_i$$

(b) Choose the following values of x:

$$x = 10^{-3},\ 1,\ 10^{2}$$

and find a method for determining a, τ_{n0}, and τ_{p0}. Show that a plot of Eq. (4-153) fits the curve of Fig. 4-24. Find E_T, using $n_i = 2.4 \times 10^{13}/\text{cm}^3$, as given by Conwell (1958)[2].

Let us now consider the nature of generation and recombination. We have mentioned in Section 3-1 that electrons can be excited to higher energies by acquiring energy in the form of heat or light. Similarly, electrons can give off energy and drop to lower levels. In equilibrium, we would expect the average energy of all the electrons to remain unchanged, and hence we postulate the *principle of detailed balance*, which states that the rates of excitation and recombination for each type of energy transformation process are the same. In a semiconductor in equilibrium, some free electrons (and the same number of free holes) are being thermally created every second, and the principle of detailed balance then tells us that an equal quantity of electron-hole pairs per second must be recombining. Now define the *generation rate* g as the additional number of pairs/cm³/sec which are created by some applied pulse of light or heat. The rate of change of electrons is then the difference between the generation and the recombination rate, so that

$$\frac{\partial n}{\partial t} = g - \frac{\Delta n}{\tau_n} \tag{4-154}$$

where Δn is the excess number of electrons, n_0 is the equilibrium number, and $n = n_0 + \Delta n$. Similarly, for holes,

$$\frac{\partial p}{\partial t} = g - \frac{\Delta p}{\tau_p} \tag{4-155}$$

PROBLEM 4-32 What is the relation between Eq. (4-155) and Eq. (4-41)?

If electrons and holes recombine directly, then $\Delta n = \Delta p$ and $\tau_n = \tau_p$. However, when traps are involved, then the two types of charges may spend different amounts of time in these traps, and the number of free electrons may not necessarily be equal to the number of free holes, so that $\tau_n \neq \tau_p$.

Let us consider now a semiconductor with N_T traps per unit volume at an energy level E_T, as shown in Fig. 4-23. In equilibrium, the definition of the Fermi function tells us that the fraction f_{T0} of traps which will be occupied is

$$f_{T0} = \frac{1}{e^{(E_T - E_F)/kT} + 1} \tag{4-156}$$

PROBLEM 4-33 Using the results of Section 3-3 and the definitions [Eq. (4-151)] of n_1 and p_1, show that

$$f_{T0} = \frac{n_0}{n_1 + n_0} = \frac{p_1}{p_0 + p_1} \tag{4-157}$$

For non-equilibrium conditions, the difference between Δp and Δn is the number of additional trapped electrons and the additional fraction is then $(\Delta p - \Delta n)/N_T$. Hence, the fraction of occupied traps is now

$$f_T = \frac{n_0}{n_1 + n_0} + \frac{\Delta p - \Delta n}{N_T} \tag{4-158}$$

Then the fraction of empty traps is $(1 - f_T)$ and we would expect the *capture rate* R_{nc} for electrons, defined as the number of electrons per second which are being trapped, to be proportional to this number and to the number of electrons, or

$$R_{nc} \propto n(1 - f_T)$$

The proportionality constant will then have dimensions of time⁻¹, and it is denoted by τ_{n0}, so

$$R_{nc} = n(1 - f_T)/\tau_{n0} \tag{4-159}$$

PROBLEM 4-34 Show that τ_{n0} is the minimum possible time constant for electron capture by combining Eq. (4-159) for $f_T = 0$ with the results of Problem 4-33.

The rate of electron emission is given by the principle of detailed balance as

$$R_{ne} = \frac{f_T\, n_1}{\tau_{n0}} \tag{4-160}$$

PROBLEM 4-35 Verify Eq. (4-160) by showing that for equilibrium, $R_{ne} = R_{nc}$.

The net rate of electron capture is then

$$R_{nc} - R_{ne} = \frac{n(1 - f_T) - n_1 f_T}{\tau_{n0}} \tag{4-161}$$

and, similarly, for holes,

$$R_{pc} - R_{pe} = \frac{p f_T - p_1(1 - f_T)}{\tau_{p0}} \tag{4-162}$$

For steady-state conditions, these net rates for electrons and holes are the same and equal to g, so that

$$\frac{n(1 - f_T) - n_1 f_T}{\tau_{n0}} = \frac{p f_T - p_1(1 - f_T)}{\tau_{p0}} \tag{4-163}$$

PROBLEM 4-36 Show that

$$g = \frac{\Delta n}{\tau_n} = \frac{\Delta p}{\tau_p} = \frac{(pn - n_i^2)}{\tau_{p0}(n_0 + n_1 + \Delta n) + \tau_{n0}(p_0 + p_1 + \Delta p)} \tag{4-164}$$

by substituting Eq. (4-158) into Eq. (4-163). [Hint: To eliminate N_T, multiply the equation involving τ_{n0} by $(p_0 + p_1 + \Delta p)$, multiply the equation involving τ_{p0} by $(n_0 + n_1 + \Delta n)$, and add.]

As the number of traps gets vanishingly small, $\Delta n = \Delta p$ and Eq. (4-164) becomes

$$\frac{1}{\tau_n} = \frac{1}{\tau_p} = \frac{n_0 + p_0 + \Delta n}{\tau_{p0}(n_0 + n_1) + \tau_{n0}(p_0 + p_1) + \Delta n(\tau_{p0} + \tau_{n0})} \tag{4-165}$$

This is of the form

$$\tau = \frac{\tau_0(1 + A\,\Delta n)}{(1 + B\,\Delta n)} \tag{4-166}$$

where A and B are constants, and it shows that the lifetime under injection conditions is different from the equilibrium lifetime.

PROBLEM 4-37

(a) Find an expression for τ_0 in terms of τ_{n0} and τ_{p0}.

(b) How does τ change as a function of Δn when $A < B$, $A > B$, and $A = B$?

(c) What does $A = B$ mean?

When $\Delta n = 0$, Eq. (4-165) reduces to Eq. (4-152), which we wished to verify.

The theory developed above has been applied to PN-junctions by Sah, Noyce, and Shockley (1957)[21], and we shall show how they predict the variation of α with current in silicon transistors. Consider, for convenience, an idealized junction, in which the traps are located at the intrinsic level ($E_T = E_I$), so that $n_1 = p_1 = n_i$. In addition, the mobilities, the lifetimes, and the minority carrier concentrations on both sides of the junction are taken to be the same. For a reverse-biased junction, the minority carrier current due to diffusion, by Eq. (4-64), is

$$J_d = \frac{2e\,D_0\,n_P}{L_0} = \frac{2e\,L_0\,n_P}{\tau_0} \tag{4-167}$$

where $n_P = p_N$, $\tau_{n0} = \tau_{p0} = \tau_0$, $D_n = D_p = D_0$, and $L_0^2 = D_0\tau_0$, by the above assumptions.

In the space-charge region of this junction, both electrons and holes are swept out by the large electric fields, so that the carrier densities are small compared to n_i. Then Eq. (4-164) becomes

$$g = -n_i/2\tau_0$$

For every pair of carriers generated, an electron must flow in the external circuit, and this is called the *recombination-generation current* J_{rg}. In a space-charge layer of width W it will be given by

$$J_{rg} = -eW\,n_i/2\tau_0$$

The ratio of the two currents is

$$J_{rg}/J_d = (n_P/4n_i)(W/L_0)$$

PROBLEM 4-38 Using the curves of resistivity vs concentration given in Sec. 8-2, show that the ratio I_{rg}/I_d is about 0.1 in germanium and about 10^3 in silicon for 2-ohm-cm material with a lifetime of 1 microsecond and a space-charge layer width of 1 micron.

Hence, the recombination-generation current due to traps in the space-charge region has a large influence on the properties of silicon devices. In previous sections, we have been assuming that the space-charge region is infinitesimally small and exerts no influence on the transistor action.

Under forward bias, the Fermi level at $x = 0$ is at the center of the forbidden gap, and hence $n = p$ at this point. Since

$$pn = n_i^2\, e^{e\mathcal{V}/kT}$$

then

$$n = p = n_i e^{e\mathcal{V}/2kT}$$

In analogy to the Debye length or the diffusion length, define a characteristic *drift length* by

$$L_{\mathcal{E}} = kT/e\mathcal{E} \tag{4-168}$$

PROBLEM 4-39

(a) Show that Eq. (4-164) reduces to

$$g = n_i\, e^{e\mathcal{V}/2kT}/2\tau_0 \tag{4-169}$$

at the center of the space-charge region.

(b) Assuming that g falls exponentially by a factor of $1/e$ in a distance $L_{\mathcal{E}}$, show by comparison with Eq. (4-167) that

$$J_{rg} = kT\, n_i\, e^{e\mathcal{V}/2kT}/\mathcal{E}\tau_0 \tag{4-170}$$

Gärtner et al. (1958)[22] have shown that the total emitter current density can now be written

$$J_E = \frac{e\, D_p\, n_{0N}}{w}\Bigg[2P - \ln(1 + P) + \frac{D_n\, n_{0P}\, w}{D_p\, p_{0N}\, L_n}(1 + P)\,P + \frac{kT\, w\sqrt{P}}{e\mathcal{E}\, D_p\, \tau_0}\Bigg] \tag{4-171}$$

PROBLEM 4-40 Verify Eq. (4-171) by adding Eqs. (4-141), (4-142), and (4-170).

The first two terms in Eq. (4-171) then give the high-level drift-diffusion current. The third term is the emitter efficiency contribution, and the fourth term is the generation-recombination current.

By using a procedure similar to that of Problem 4-28, it is found that the current amplification depends upon the emitter current density in accordance with

$$1 - \alpha = \frac{A\, f(J_E) + B\, g(J_E) + h(J_E)}{(1 + 2P)\sqrt{P} + C(1 + P)} \tag{4-172}$$

where

$$A = \frac{w^2}{2L_p^2} + \frac{sw}{2D_p}$$

$$B = D_n\, n_{0P}\, w/D_p\, p_{0N}\, L_n$$

$$C = kTw/2e\mathcal{E}\, D_p\, \tau_0$$

$$f(J_E) = (1 + P)\sqrt{P}$$

$$g(J_E) = (1 + P)(1 + 2P)\sqrt{P}$$

$$h(J_E) = C(1 + P)$$

PROBLEM 4-41 Verify Eq. (4-172).

A plot of $(1 - \alpha)$ vs J_E as measured on a germanium *pnp* transistor by Gärtner et al. (1958) is shown in Fig. 4-25. They have been able to fit the theoretical expression assuming reasonable values for the various parameters. The explanation for the rise and the subsequent fall in alpha is as follows: at very low currents, the recombination-generation process is the dominant one and thus detracts from the emitter efficiency. As the current rises, the diffusion contribution increases, so that γ goes up. At still higher levels, the field begins to increase the effective diffusion length. At quite high levels, the large number of injected carriers decreases the base resistivity and γ begins to fall again, accounting for the minimum in the curve of Fig. 4-25.

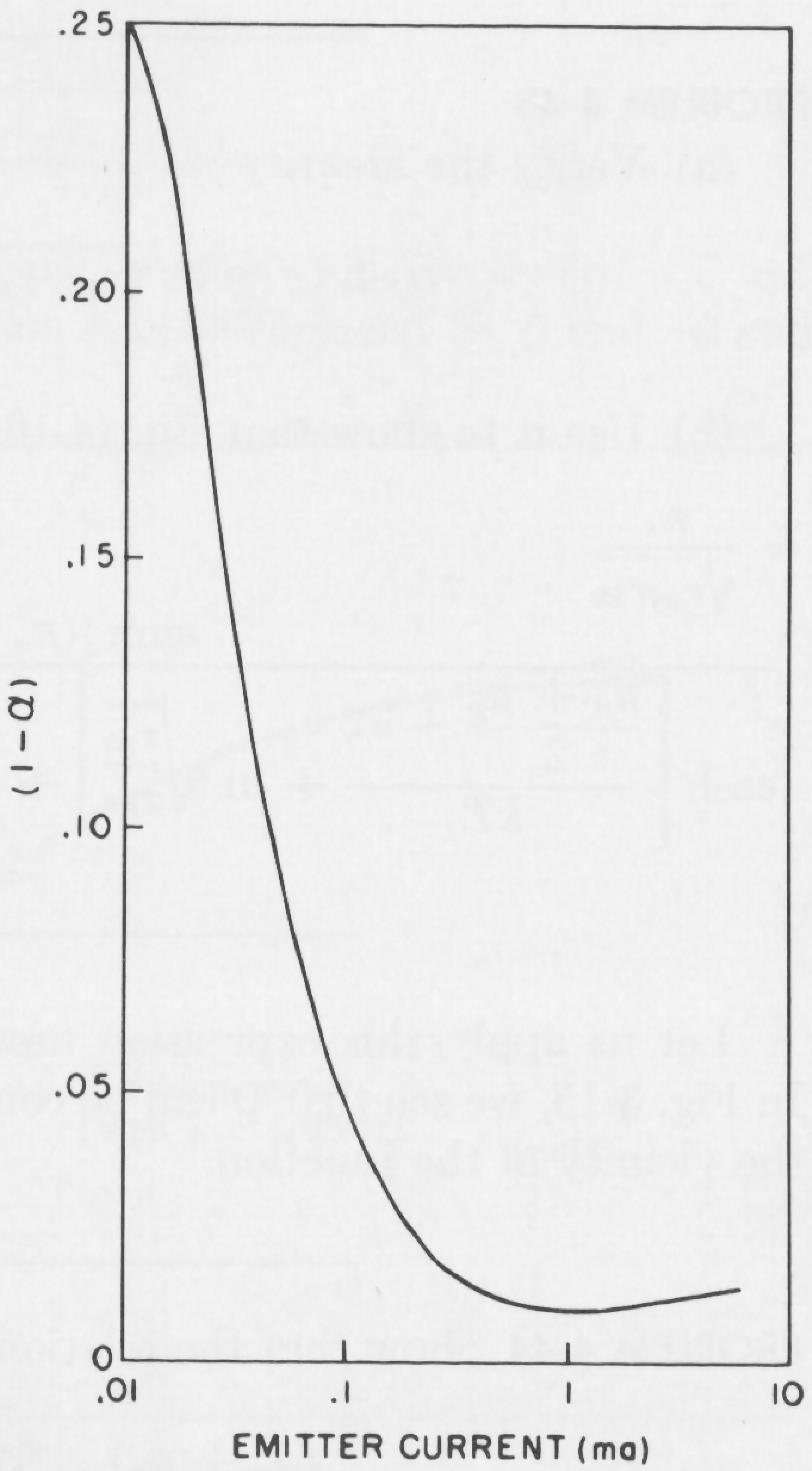


FIG. 4-25. The dependence of amplification on current density in a junction transistor, taking into account generation-recombination processes in the space charge layer. (From W. W. Gärtner, et al, Proc. IRE **46,** 1875, 1950, with permission.)

To consider a more detailed calculation of J_{rg}, we use the definition given by Eq. (1-22) for current-density

$$J_{rg} = nev = e\int g\, dx \tag{4-173}$$

PROBLEM 4-42 Show the plausibility of Eq. (4-173).

An expression for g comes by combining the following:

$$p_1 = n_i\, e^{(E_I - E_T)/kT} \qquad n_1 = n_i\, e^{(E_T - E_I)/kT}$$

$$p = n_i\, e^{-(\bar{\mu}_p + e\mathcal{V})/kT} \qquad n = n_i\, e^{(\bar{\mu}_n + e\mathcal{V})/kT} \qquad \text{(4-33), (4-34)}$$

$$E_I = [E_C + E_V + kT \ln (N_C/N_V)]/2 = -e\mathcal{V} \qquad \text{(3-26), (3-31)}$$

with Eq. (4-164).

PROBLEM 4-43

(a) Verify the identity

$$\cosh\,(x + \ln \sqrt{\tau_{p0}/\tau_{n0}}) = \frac{\tau_{p0}\, e^{x} + \tau_{n0}\, e^{-x}}{2\sqrt{\tau_{p0}\,\tau_{n0}}}$$

(b) Use it to show that Eq. (4-164) becomes

$$g = \frac{n_i}{\sqrt{\tau_{p0}\tau_{n0}}} \times \frac{\sinh\,[(\bar{\mu}_n - \bar{\mu}_p)/2kT]}{\cosh\left[\dfrac{\dfrac{\bar{\mu}_n + \bar{\mu}_p}{2} + e\mathcal{V}}{kT} + \ln\sqrt{\dfrac{\tau_{p0}}{\tau_{n0}}}\right] + e^{(\bar{\mu}_p - \bar{\mu}_n)/2kT} \cosh\left[\dfrac{E_T - E_I}{kT} + \ln\sqrt{\dfrac{\tau_{p0}}{\tau_{n0}}}\right]} \qquad \text{(4-174)}$$

Let us apply this expression first for the case of small forward bias. In Fig. 3-15, we see that $\mathcal{V}$ can be considered to be a linear function of x in the vicinity of the junction.

PROBLEM 4-44 Show that the relation

$$e\mathcal{V} + \frac{(\bar{\mu}_p + \bar{\mu}_n)}{2} = -(\mathcal{V}_0 + \bar{\mu}_p - \bar{\mu}_n)\, x/W \qquad \text{(4-175)}$$

satisfies the following requirements:

(a) At $x = 0$, $-e\mathcal{V}$ falls half-way between $\bar{\mu}_n$ and $\bar{\mu}_p$.
(b) At $x = \pm W/2$, $-e\mathcal{V}$ is increased or decreased to $\mathcal{V}_B$ of Fig. 4-6a,

PROBLEM 4-45 By letting

$$z = \sqrt{\tau_{p0}/\tau_{n0}}\, e^{-(\mathcal{V}_0 + \bar{\mu}_p - \bar{\mu}_n)x/W}$$

show that Eq. (4-173) becomes

$$J_{rg} = \frac{en_i}{\sqrt{\tau_{p0}\,\tau_{n0}}} \frac{2Wf(b)\sinh\,[(\bar{\mu}_n - \bar{\mu}_p)/2kT]}{(e\mathcal{V}_0 - \bar{\mu}_n + \bar{\mu}_p)/kT} \qquad \text{(4-176)}$$

where $$f(b) = \int_{z_1}^{z_2} \frac{dz}{z^2 + 2bz + 1}$$

$$b = e^{(\bar{\mu}_p - \bar{\mu}_n)/2kT} \cosh\left(\frac{E_T - E_I}{kT} + \ln\sqrt{\tau_{p0}/\tau_{n0}}\right)$$

$$z_{1,2} = \sqrt{\tau_{p0}/\tau_{n0}}\, e^{\mp[e\mathcal{V}_0 - (\bar{\mu}_n - \bar{\mu}_p)]/2kT}$$

If the applied voltages as measured by $(\bar{\mu}_n - \bar{\mu}_p)$ are several kT less than $e\mathcal{V}_0$, then the limits of integration can be replaced by 0 and ∞ and

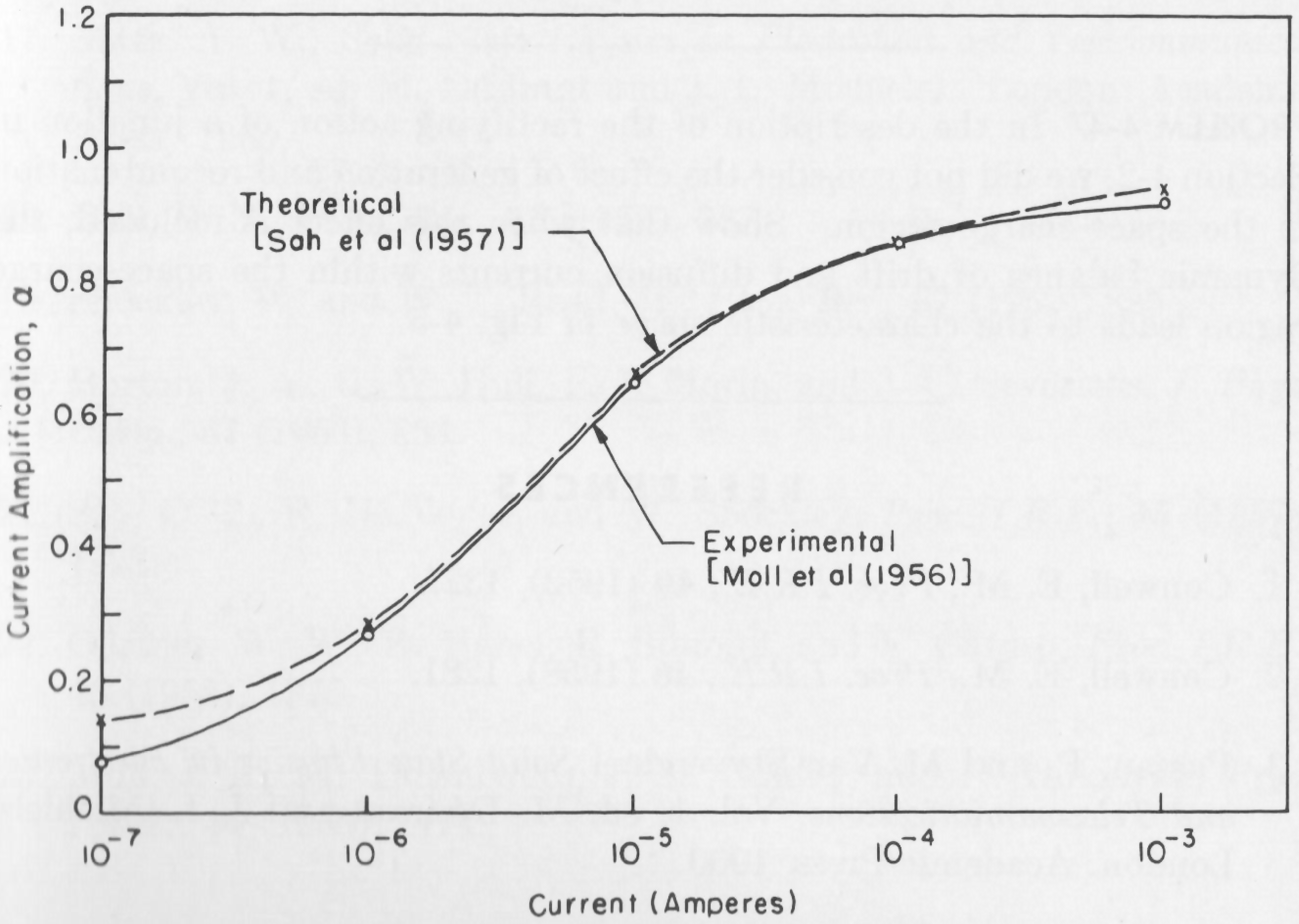


FIG. 4-26. Current amplification as a function of current density according to the theory of Sah, Noyce and Shockley. (From Sah, et al, Proc. IRE **45,** 1228, 1957, and J. L. Moll, et al, Proc. IRE **44,** 1174, 1956, with permission.)

$f(b)$ can be evaluated. Standard tables show $f(0) = \pi/2$ and that $f(b)$ monotonically decreases with b.

PROBLEM 4-46 Show that Eq. (4-81) can be written

$$\alpha = \frac{\operatorname{sech} w/L_p}{1 + \dfrac{J_{rg} + J_{En}}{J_{Ep}} \tanh w/L_p} \qquad (4\text{-}177)$$

By using J_{rg} as given by Eq. (4-176), and J_{En} and J_{Ep} from Section 4-5, the current amplification α was calculated for $w/L_n = 0.1$, with the result shown in Fig. 4-26. For comparison, there are also shown the characteristics of a silicon transistor, as measured by Moll et al. (1956)[23].

In the development given above, we have considered only the variation of α with J_E. Sah, Noyce, and Shockley (1957)[21] also show how the generation-recombination current affects the ideal rectifier characteristic of Fig. 4-5, and the agreement of theory and experiment is shown in this paper.

PROBLEM 4-47 In the description of the rectifying action of a junction in Section 4-2, we did not consider the effect of generation and recombination in the space-charge region. Show that when this effect is included, the dynamic balance of drift and diffusion currents within the space-charge region leads to the characteristic curve of Fig. 4-5.

REFERENCES

1. Conwell, E. M., *Proc. I.R.E.*, **40** (1952), 1327.
2. Conwell, E. M., *Proc. I.R.E.*, **46** (1958), 1281.
3. Passau, P. and M. Van Styvendael *Solid State Physics in Electronics and Telecommunications*, Vol. 1, ed. M. Désirant and J. L. Michiels. London: Academic Press, 1960.
4. Muss, D. P., *J. Appl. Phys.*, **26** (1955), 1514.
5. Shockley, W., *Electrons and Holes in Semiconductors*. Princeton: D. Van Nostrand, 1950.

6. Domenicali, C. A., *Rev. Mod. Phys.*, **26** (1954), 237.
7. Nussbaum, A., *Proc. I.R.E.*, **43** (1955) 315.
8. Armstrong, L. D., C. L. Carlson, and M. Bentivegna, *RCA Rev.*, **17** (1956), 37.
9. Long, D., *J. Appl. Phys.*, **28** (1957) 1219.
10. Low, G. E. E., *Proc. Phys. Soc.*, **B68** (1955), 310.
11. Granville, J. W., *J. Elec.*, **1** (1956), 565.
12. Hall, R. N., *Proc. I.R.E.*, **40** (1952), 1512.
13. Long, D., Honeywell Research Center Report No. GR 4599-R1 (1956), unpublished.
14. Webster, W. M., *Proc. I.R.E.*, **42** (1954), 914.
15. Rittner, E. S., *Phys. Rev.*, **94** (1954), 1161.
16. Misawa, T., *J. Phys. Soc. Japan*, **11** (1956), 728.
17. Matz, A. W., *Solid State Physics in Electronics and Telecommunications*, Vol. 1, ed. M. Désirant and J. L. Michiels. London: Academic Press, 1960.
18. Hall, R. N., *Phys. Rev.*, **87** (1952), 387.
19. Shockley, W. and W. T. Read, Jr., *Phys. Rev.*, **87** (1952), 835.
20. Burton, J. A., G. W. Hull, F. J. Morin, and J. C. Severiens, *J. Phys. Chem.*, **57** (1953), 853.
21. Sah, C.-T., R. N. Noyce, and W. Shockley, *Proc. I.R.E.*, **45** (1957), 1228.
22. Gärtner, W. W., R. Hanel, R. Stampfl, and F. Caruso, *Proc. I.R.E.*, **46** (1958), 1875.
23. Moll, J. L., M. Tanenbaum, J. M. Goldey, and N. Holonyak, *Proc. I.R.E.*, **44** (1956), 1174.

GENERAL REFERENCES

Gärtner, W. W., *Transistors: Principles, Design, and Applications*. Princeton: D. Van Nostrand, 1960.

Greiner, R. A., *Semiconductor Devices and Applications.* New York: McGraw-Hill Book Co. 1961.

These two recent books give a more cursory introduction to semiconductor physics than that of Chapters 1–3, but they cover circuitry extensively and present quite a bit of material on specialized devices and effects.

5

PERIODIC STRUCTURES AND ENERGY BANDS

5-1 Symmetry and Lattices

Ordinarily, when we think of *symmetry*, we have in mind bilateral symmetry. The human body roughly illustrates bilateral symmetry, as shown in Fig. 5-1a. Another type of symmetry found in nature is *rotational symmetry*, as shown by the starfish of Fig. 5-1b. The two species in this figure illustrate *five-fold* and *eleven-fold* symmetry, and we shall soon see that these particular types of rotational symmetry are not permitted in solid state physics. In Chapter 1, we defined a *crystal* as a regular array of atoms. This means that a crystal possesses *translational symmetry;* that is, atoms or groups of atoms are repeated at fixed intervals called the *lattice constants.* Figures 1-19 and 1-20 show one-dimensional translational symmetry, with the lattice constant denoted as L.

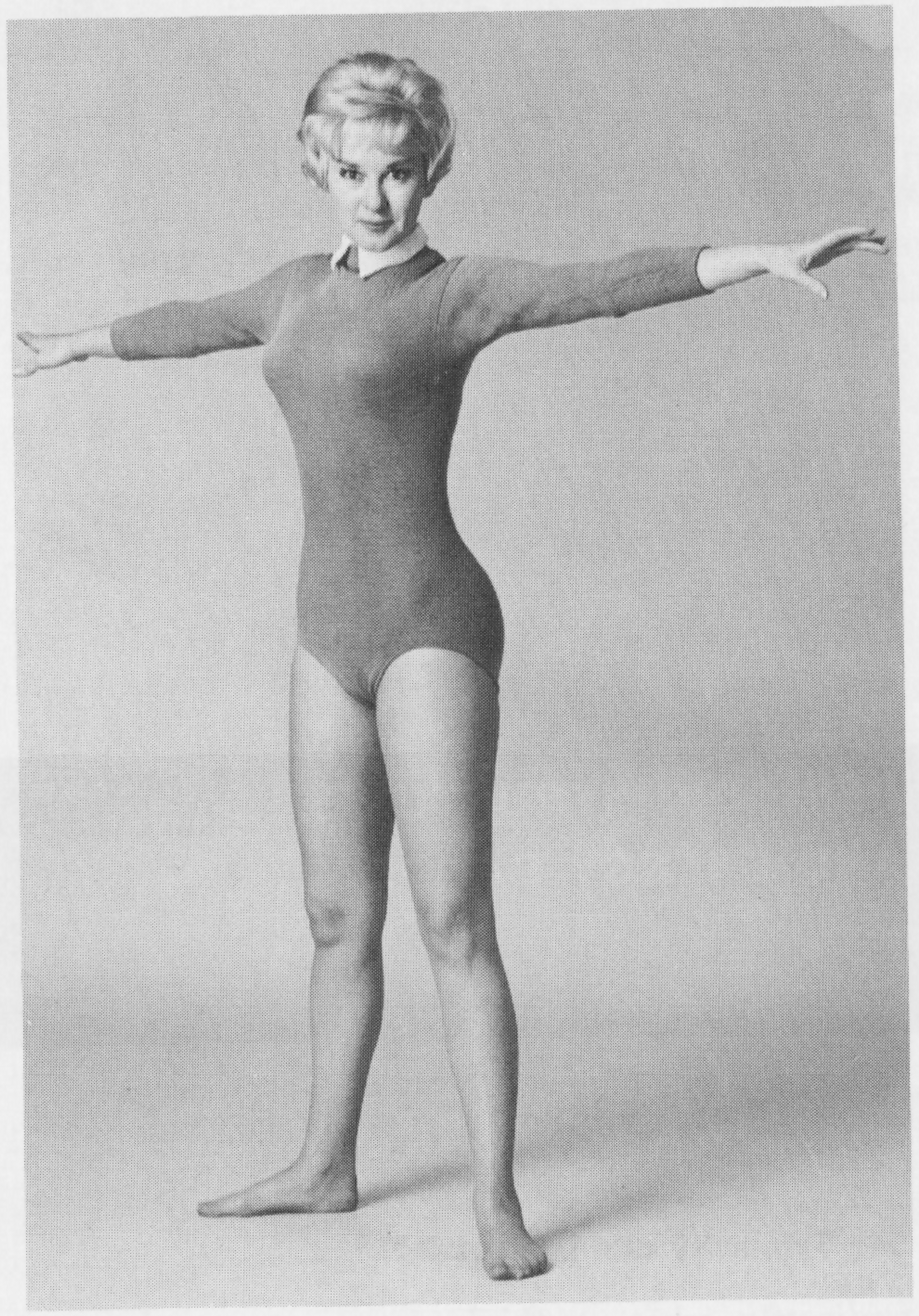
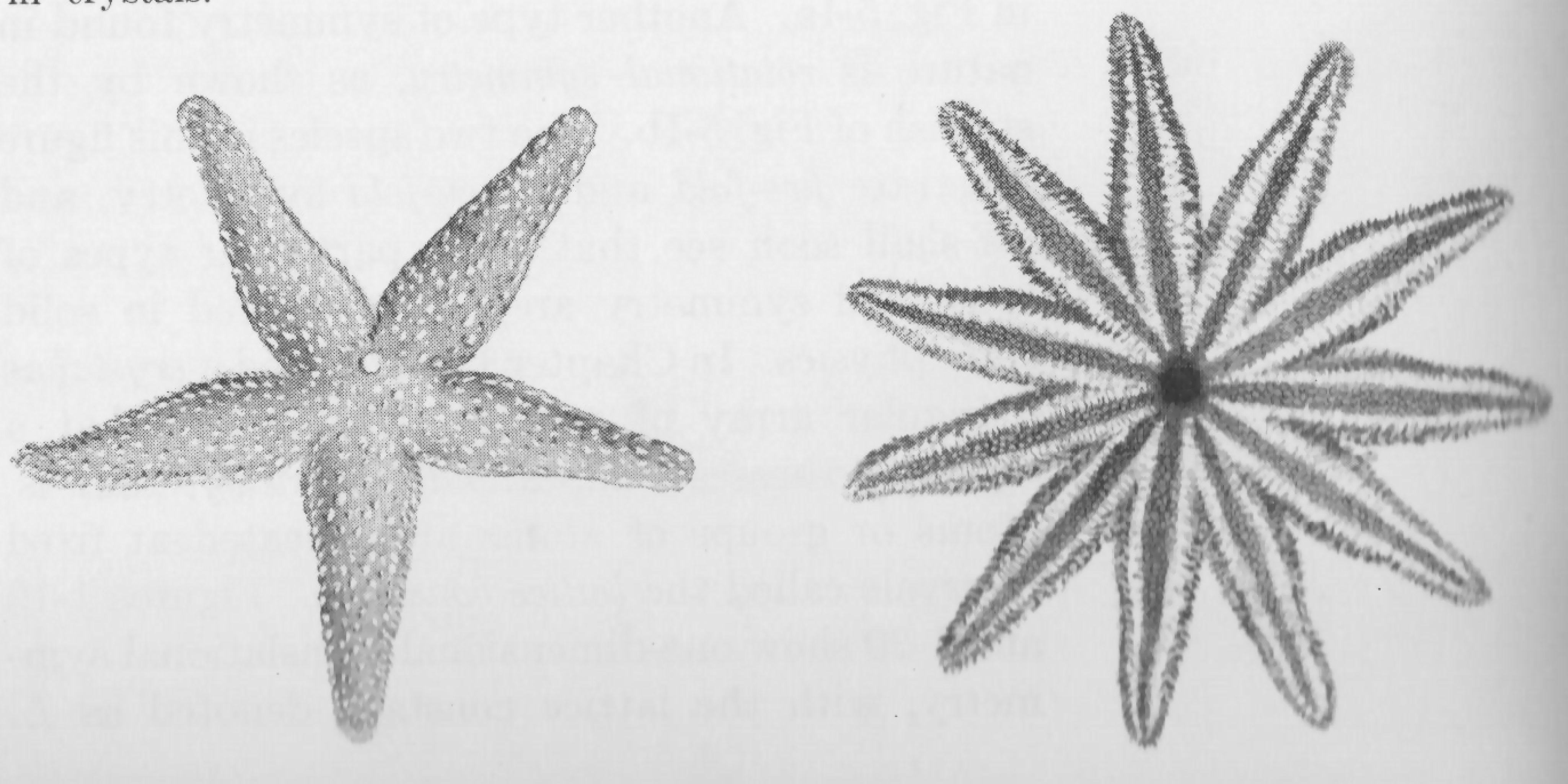


FIG. 5-1 (a). An example of bilateral symmetry.

FIG. 5-1 (b). Rotational symmetry found in living objects but not in crystals.



It might be expected intuitively that there would be an infinite number of ways of arranging atoms to form crystals, but there are only 230 (disregarding differences in chemical composition). We shall not prove this rigorously, but we shall indicate how it may be demonstrated.

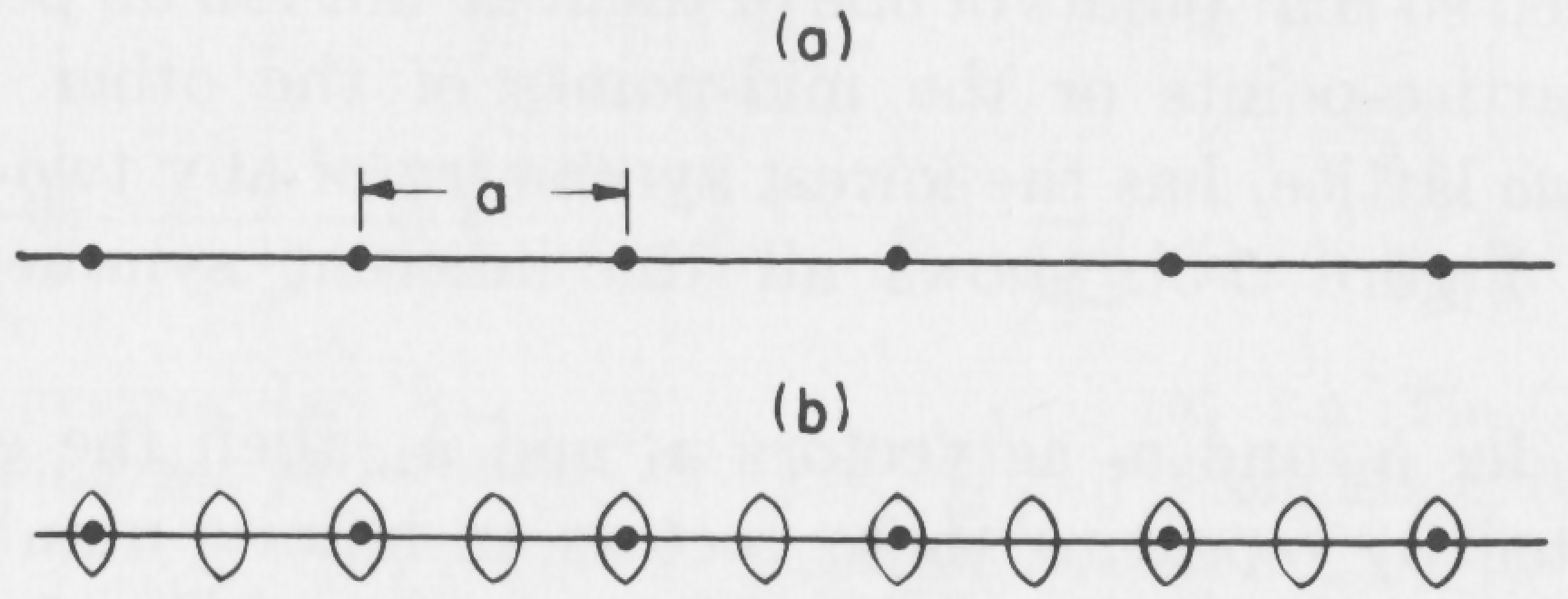


FIG. 5-2. A one-dimensional lattice. (From W. L. Bond, University of Toronto Studies, Geological Series, No. 51, 1947, with permission.)

Let us start with the infinite row of equally-spaced points in Fig. 5-2a. This row is seen to have the following symmetry properties:

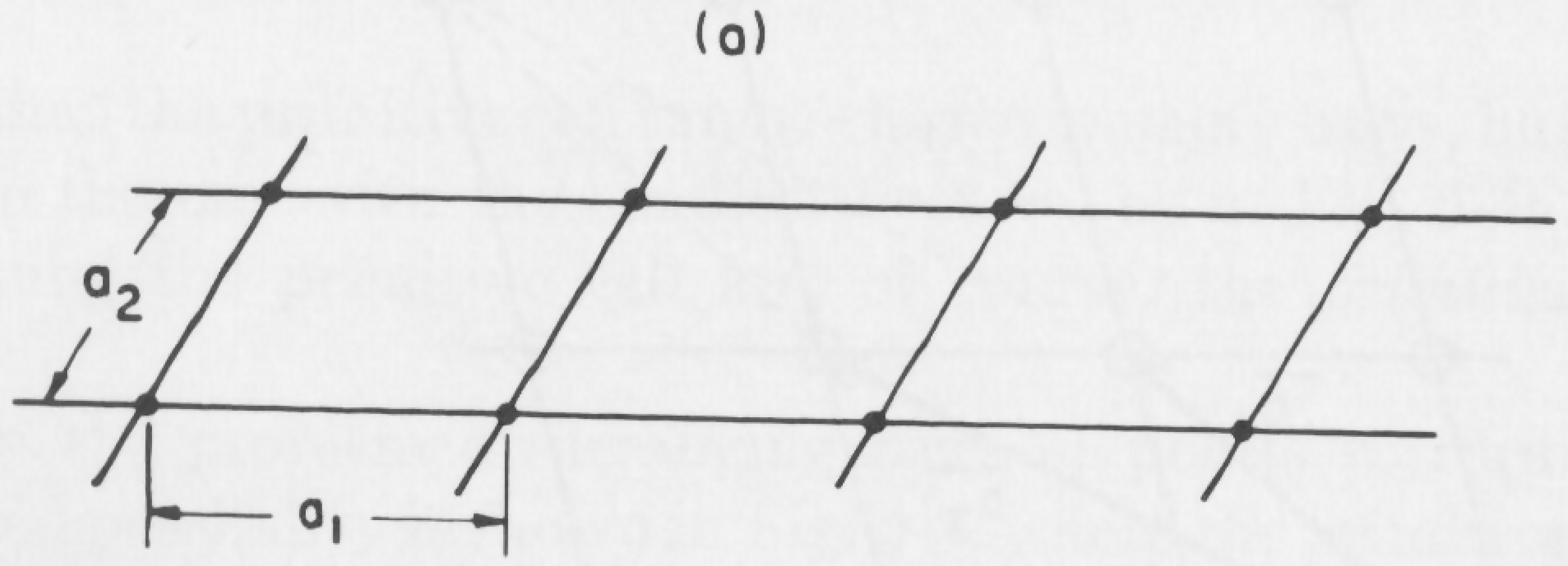
(1) Translational symmetry, with the lattice constant a.

(2) Each point can act as a mirror for all the other points, and the result will be a new row which looks exactly like the original.

(3) A point halfway between any two adjacent points can also act as a mirror.

(4) A line perpendicular to the plane of the paper and passing through any point can act as a *two-fold rotation axis,* so that a rotation of 180 deg brings the row into coincidence with itself and a second 180-deg rotation returns it to its original position.

(5) Two-fold axes, marked with a boat-shaped symbol, also exist halfway between each point (an equilateral triangle will be used to denote three-fold axes, a square for four-fold axes, and so on).



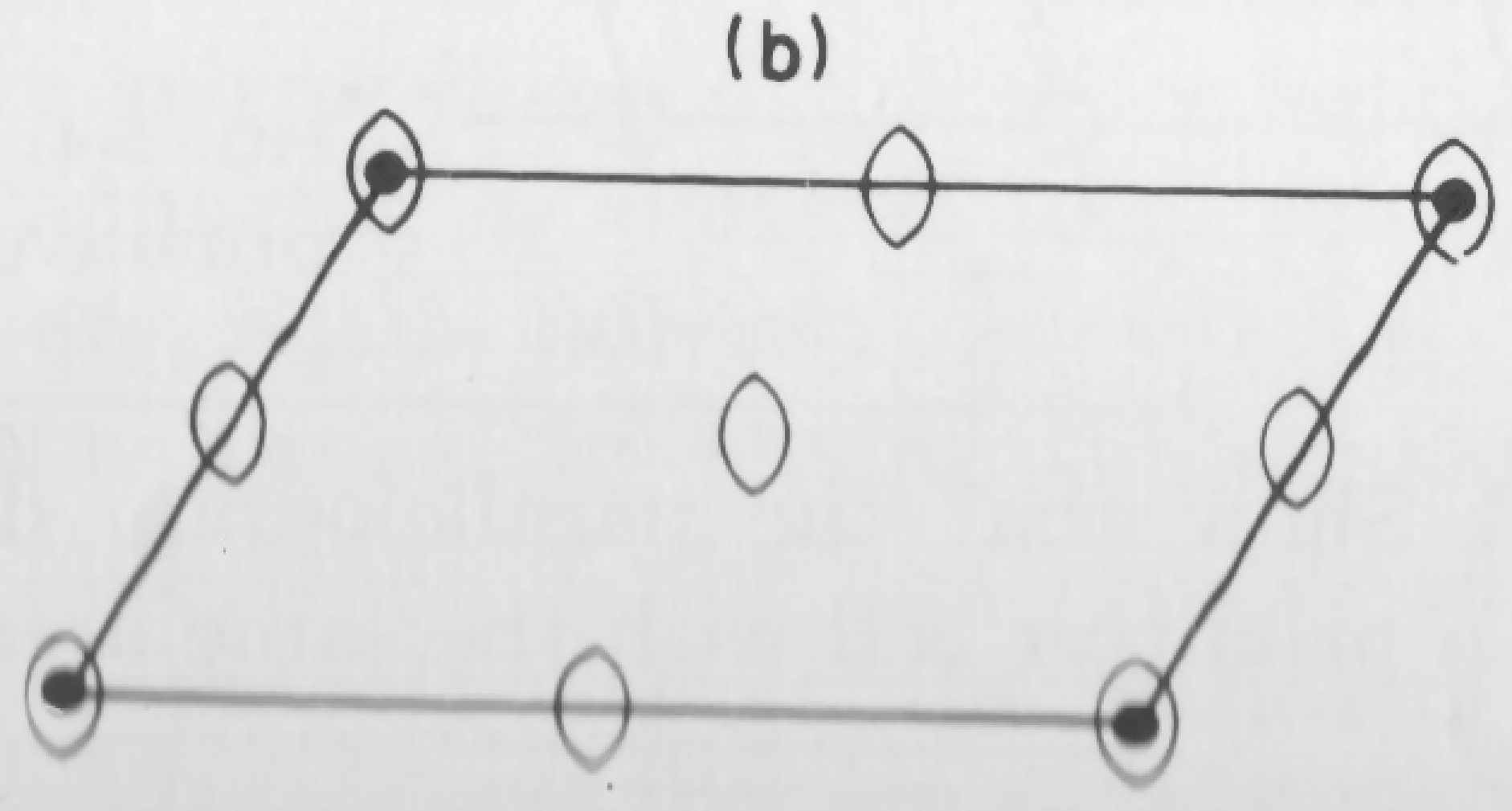


FIG. 5-3 (a). The oblique lattice.
FIG. 5-3 (b). The inherent symmetry of the oblique lattice. (From Bond, *op. cit.*)

We can now combine two or more rows to form *plane lattices*, as shown in Fig. 5-3, following a development of Bond (1947)[1]. The figure shows two lattice constants, a_1 and a_2, as are required for a two-dimensional arrangement. We have chosen a_2 to be different from a_1, and the two rows have been placed so that points of one of them do not fall on perpendiculars through the lattice-points or the mid-points of the other. This type, called an *oblique* lattice, has the lowest symmetry of any two-dimensional arrangement. Figure 5-3b shows all the inherent symmetries of the oblique lattice.

If we consider a_1 and a_2 as vectors $\mathbf{a}_1$ and $\mathbf{a}_2$, then the entire lattice can be generated by repeating these vectors an infinite number of times. In Fig. 5-4a, we have reproduced the oblique lattice of Fig. 5-3 and drawn the lattice points as circles of finite area. The parallelogram determined by $\mathbf{a}_1$ and $\mathbf{a}_2$ forms a *primitive cell;* this definition comes from the fact that the total area of the sections of the four lattice points enclosed by a single parallelogram equals one complete point.

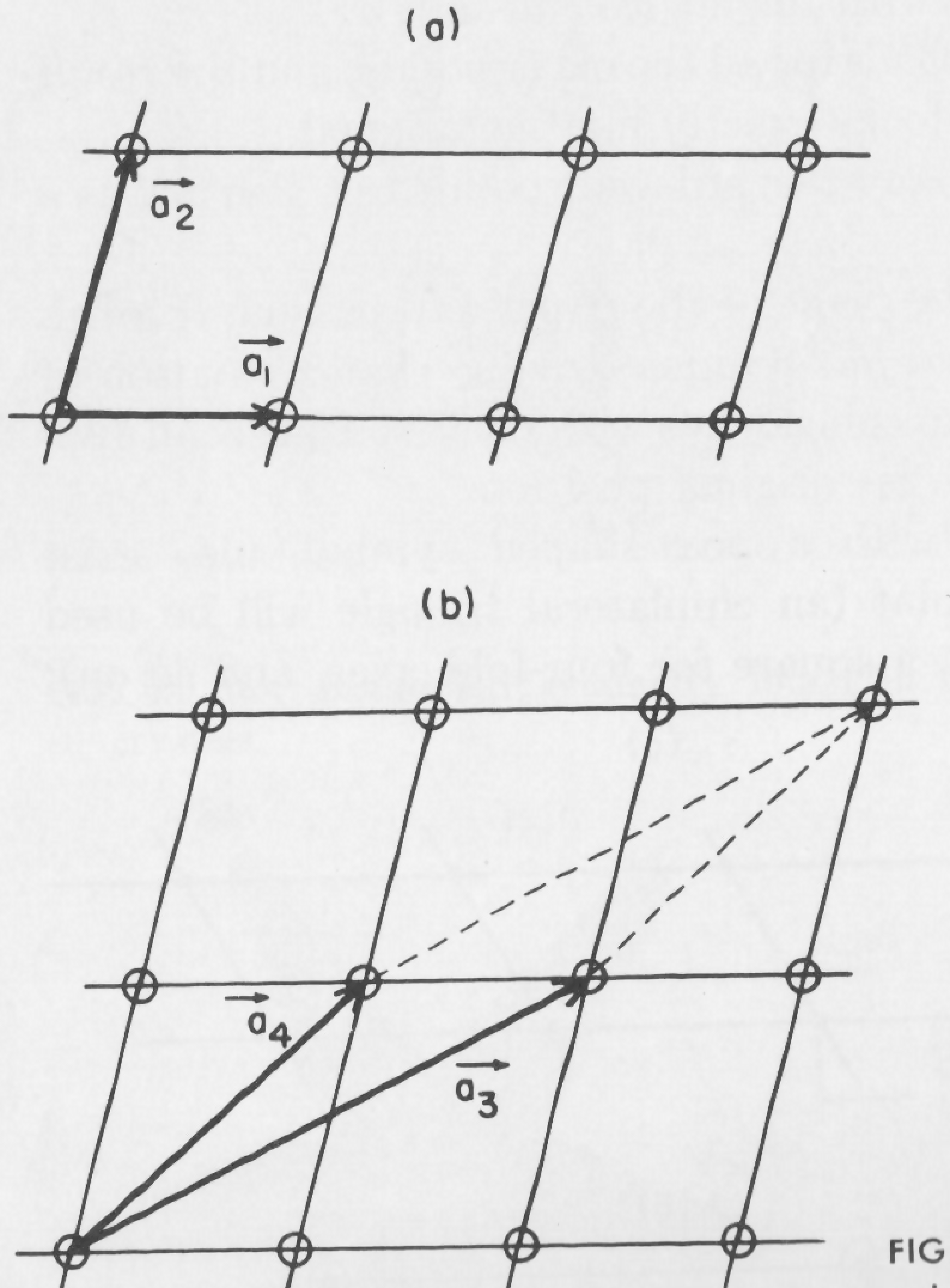


FIG. 5-4. Two possible choices for the primitive vectors of the oblique lattice.

PROBLEM 5-1 Show that the parallelogram determined by $\mathbf{a}_3$ and $\mathbf{a}_4$ in Fig. 5-4b is a primitive cell with the same area as that in Fig. 5-4a.

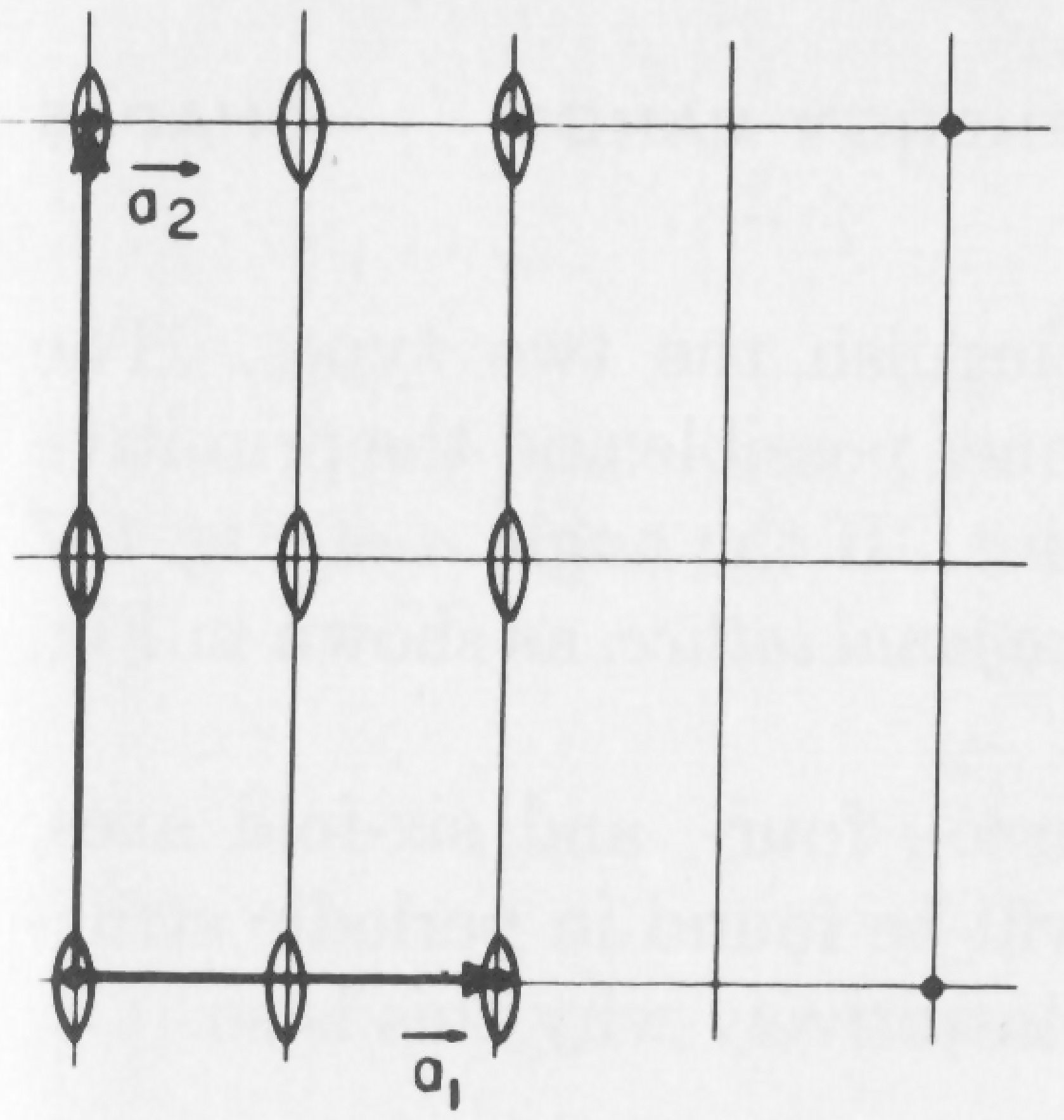


FIG. 5-5. The rectangular lattice. (From Bond, *op. cit.*)

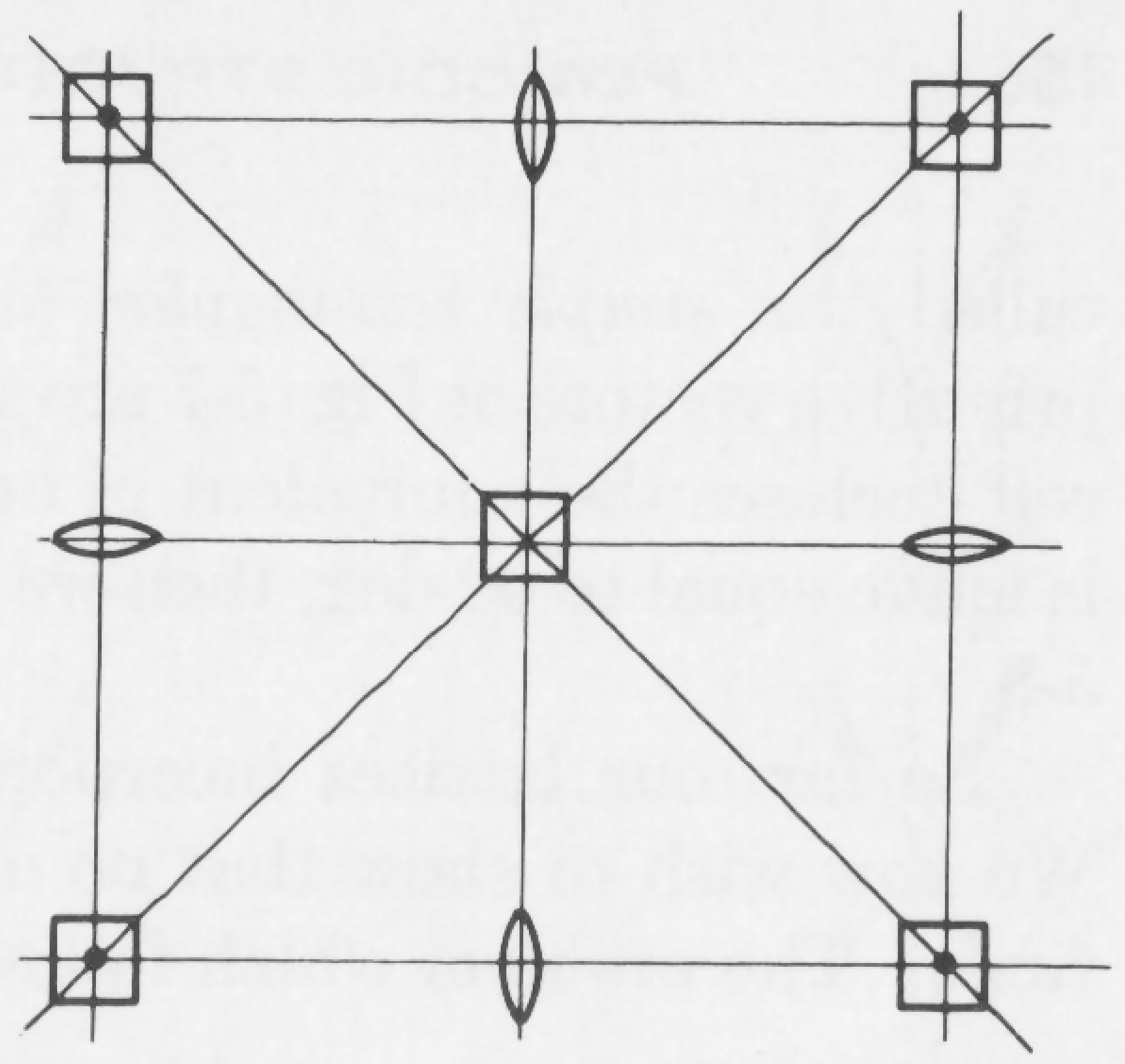


FIG. 5-6. The square lattice. (From Bond, *op. cit.*)

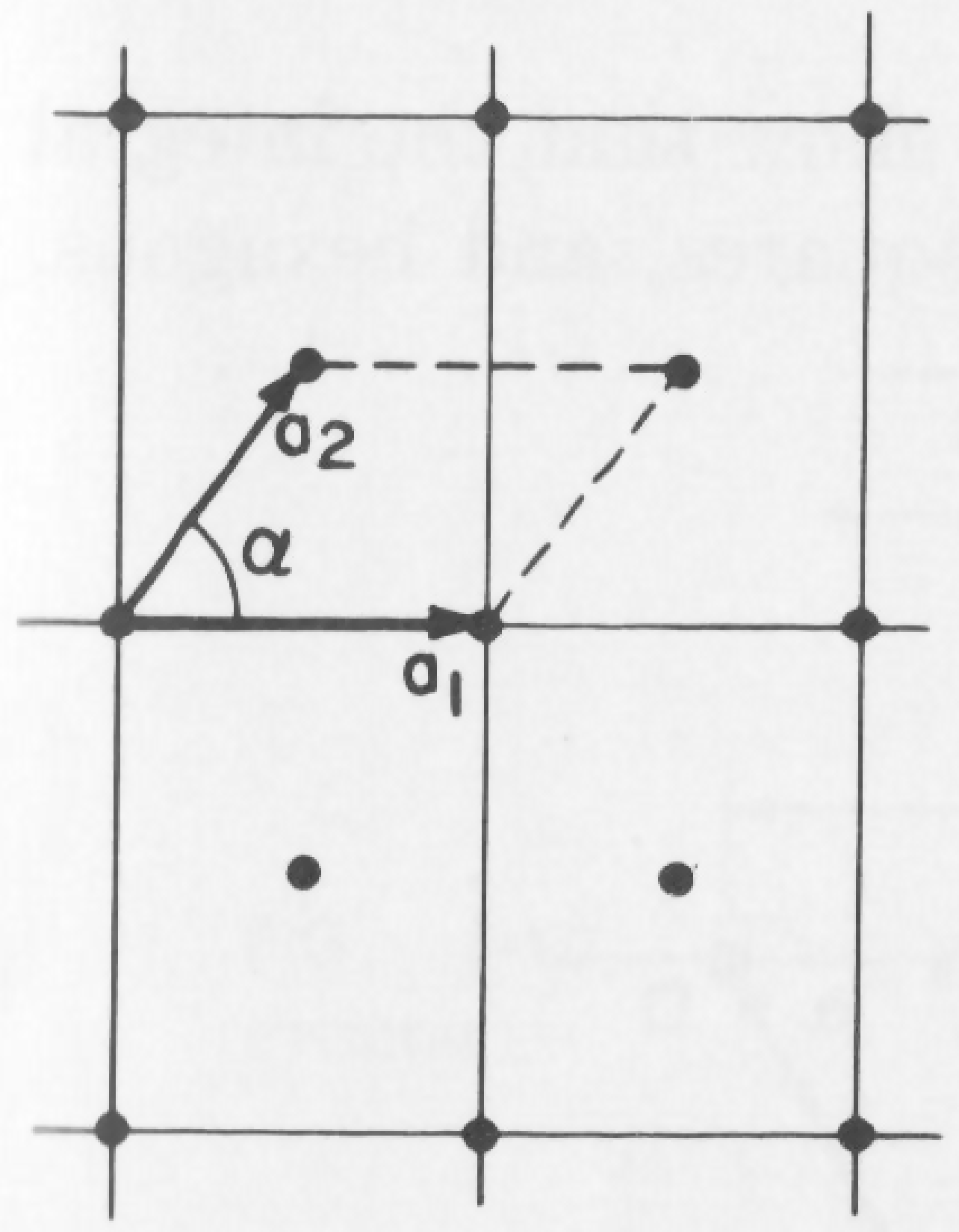


FIG. 5-7. The centered rectangular lattice.

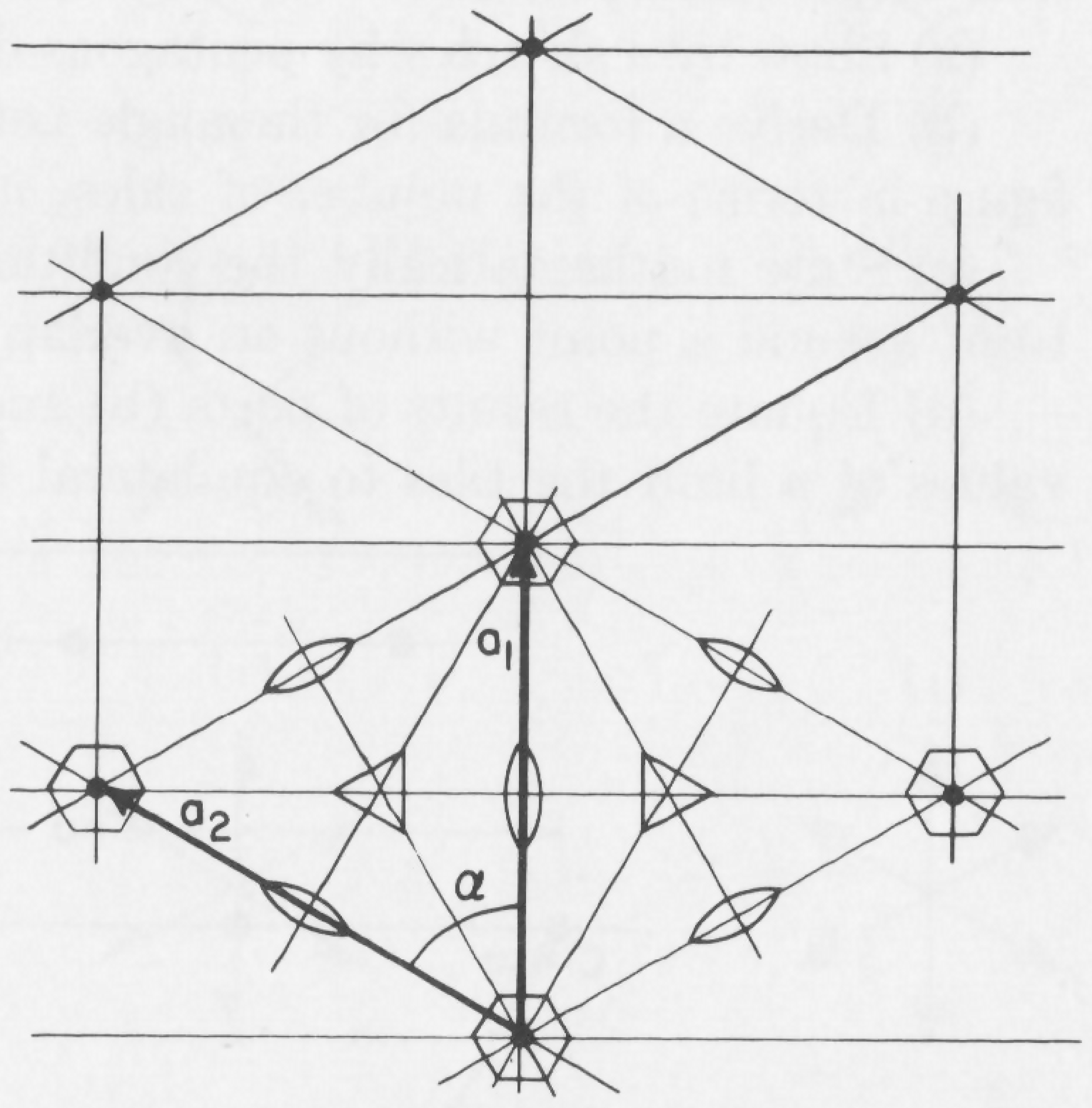


FIG. 5-8. The hexagonal lattice. (From Bond, *op. cit.*)

Figure 5-4 shows that the primitive cell can be chosen in many ways, but we shall generally use the one with the smallest vectors, as in Fig. 5-4a. The vectors determining the primitive cell are, of course, the *primitive lattice vectors.*

We return now to the problem of arranging rows of points to form plane lattices. A second possibility is shown in Fig. 5-5, where the primitive vectors now form a right angle. In addition to the two-fold axes found in the oblique lattice, we have mirror-lines, denoted by solid lines. This lattice is *rectangular*, and by making $a_1 = a_2$, we form the *square* lattice of Fig. 5-6. The solid lines are, again, mirrors.

If the rectangular lattice is formed by having the second row lie on perpendiculars through the mid-points of the first row, as shown in Fig. 5-7, we obtain the *centered rectangular lattice;* the lattice of Fig. 5-5 is called the *simple rectangular lattice* to distinguish the two types. The primitive vectors of Fig. 5-7 are the shortest ones possible and the primitive cell encloses the equivalent of one lattice point. If the angle α of Fig. 5-7 is made equal to 60 deg, then we have the *hexagonal lattice*, as shown in Fig. 5-8.

So far, our lattices incorporate two-, three-, four-, and six-fold axes. We now wish to show that no other types will be found in periodic structures. The problem which follows indicates intuitively why this is so.

PROBLEM 5-2 In laying tile floors, using regular figures, it has been found that only squares, equilateral triangles, and hexagons can be used.

(a) Show by a sketch why pentagons do not fit together.

(b) Derive a formula for the angle between adjacent sides of a regular figure in terms of the number of sides, n.

(c) State mathematically the condition for a group of regular figures to fit around a point without an overlap or a gap.

(d) Equate the results of parts (b) and (c), and show that the integral values of n limit the tiles to equilateral triangles, squares, and hexagons.

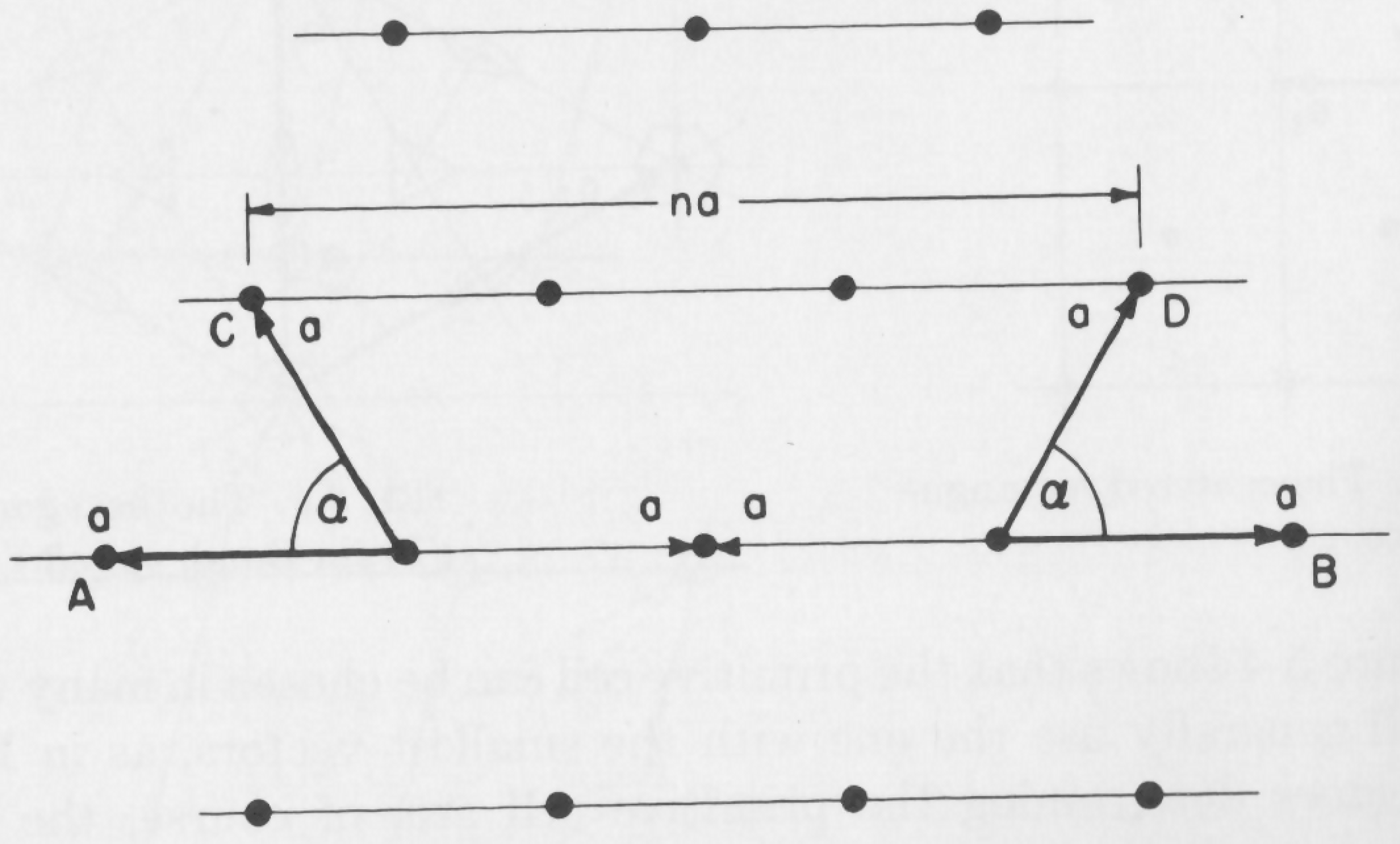


FIG. 5-9. A hexagonal lattice used to illustrate the relation between translational and rotational symmetries. (After L. V. Azaroff, *Introduction to Solids*, McGraw-Hill, New York, 1960, with permission.)

Consider now as an example the hexagonal lattice of Fig. 5-9. All the vectors in the figure have a length a, and the angle α for a six-fold axis is 60 deg. The horizontal vector terminating at A will terminate at C after a clockwise rotation through an angle α, and, similarly, the one at B will go to D after an equal counter-clockwise rotation, both rotations being permitted by the symmetry. Points C and D must be lattice points, and hence their separation will be an integral multiple, na, of a. Also, we have shown the centers of rotation as $2a$ apart, but this distance can be any integral multiple, ma, of a. Hence,

$$na = ma + 2a \cos\alpha$$

or

$$\cos\alpha = N/2 \tag{5-1}$$

where $N = n - m$ is also an integer. Equation (5-1) can have solutions as shown in Table 5-1.

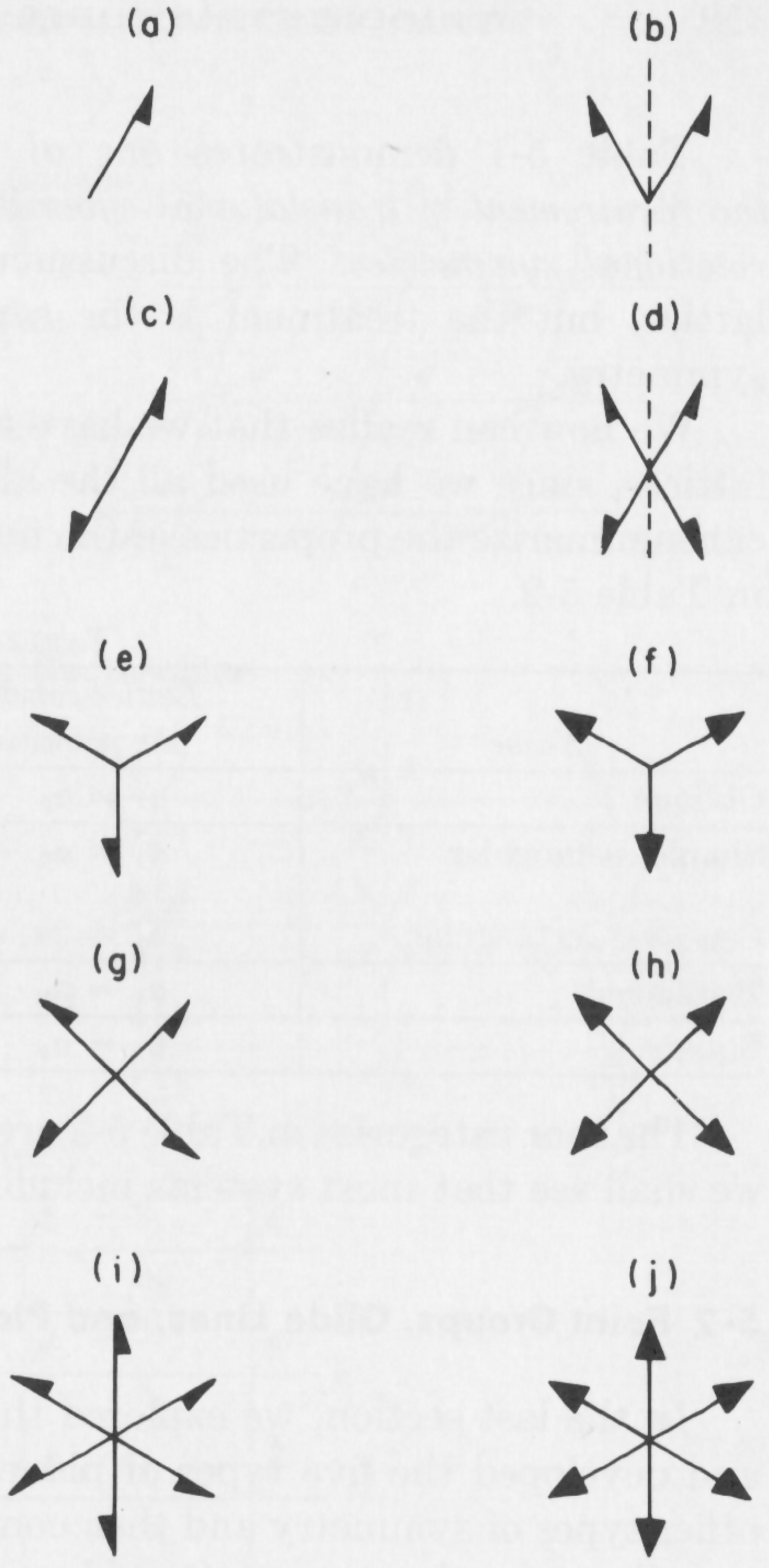


FIG. 5-10. The plane point groups.

TABLE 5-1

N	$\cos\alpha$	α	Axis symmetry
-2	-1	180°	two-fold
-1	$-\frac{1}{2}$	120°	three-fold
0	0	90°	four-fold
1	$\frac{1}{2}$	60°	six-fold
2	1	0°	one-fold

Table 5-1 demonstrates one of the basic properties of a lattice: *the requirement of translational symmetry permits only a limited number of rotational symmetries.* The discussion above was based on a hexagonal lattice, but the treatment is the same for any lattice with rotational symmetry.

We now can realize that we have exhausted all possible types of plane lattices, since we have used all the allowable rotational symmetries. We can summarize the properties of the five lattices we have derived, as shown in Table 5-2.

TABLE 5-2

Name	*Lattice constants of the primitive cell*	*Lattice angle of the primitive cell*
Oblique	$a_1 \neq a_2$	$\alpha \neq 90°$
Simple rectangular	$a_1 \neq a_2$	$\alpha = 90°$
Centered rectangular	$a_1 = 2a_2 \cos \alpha$	$\alpha \neq 60°$
Hexagonal	$a_1 = a_2$	$\alpha = 60°$
Square	$a_1 = a_2$	$\alpha = 90°$

The four categories in Table 5-2 are called *systems.* In three dimensions, we shall see that most systems include several types of lattice.

5-2 Point Groups, Glide Lines, and Plane Groups

In the last section, we explored the concept of translational symmetry and developed the five types of plane lattices. Let us now take up some other types of symmetry and then combine all of the various kinds to form two-dimensional patterns. Consider first, as an example of an unsymmetrical element, a half-arrow, as shown in Fig. 5-10a. This arrow can be combined with another just like it to produce a symmetrical pattern about a point in two ways: by reflection in a mirror (Fig. 5-10b), or by a two-fold rotation (Fig. 5-10c). The two-fold rotation can lead to a higher symmetry by combining it with a mirror operation (Fig. 5-10d), and further possible arrangements are shown in Figs. 5-10e through 5-10j. In these figures we have included only the rotational symmetries which were shown in the previous section to be consistent with a plane lattice. Hence, we have ten arrangements possible around a point, which are called *point groups.*

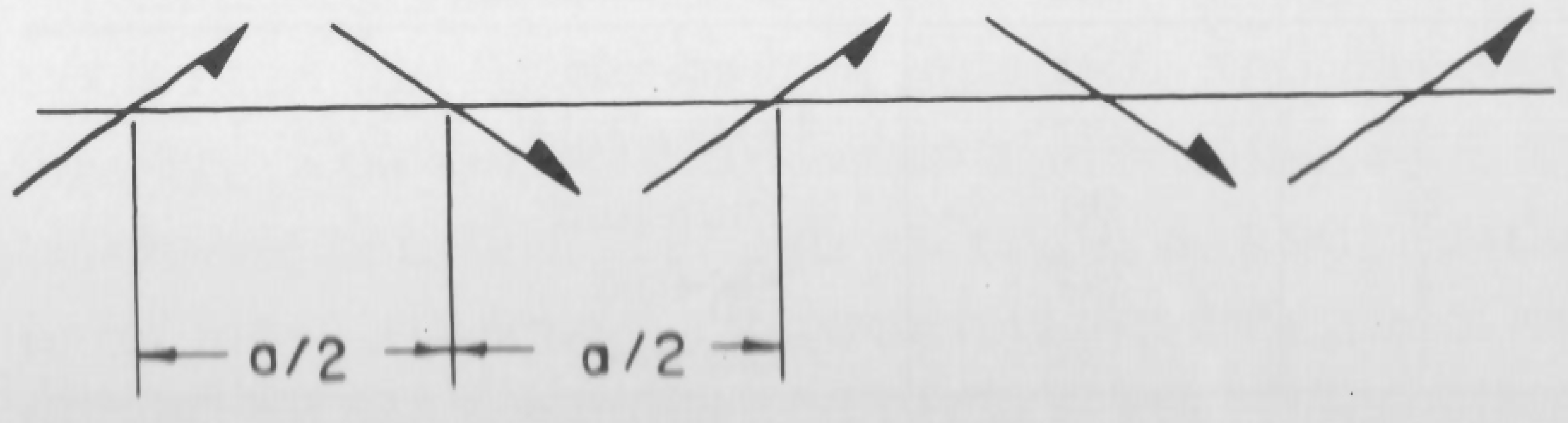


FIG. 5-11. The glide line.

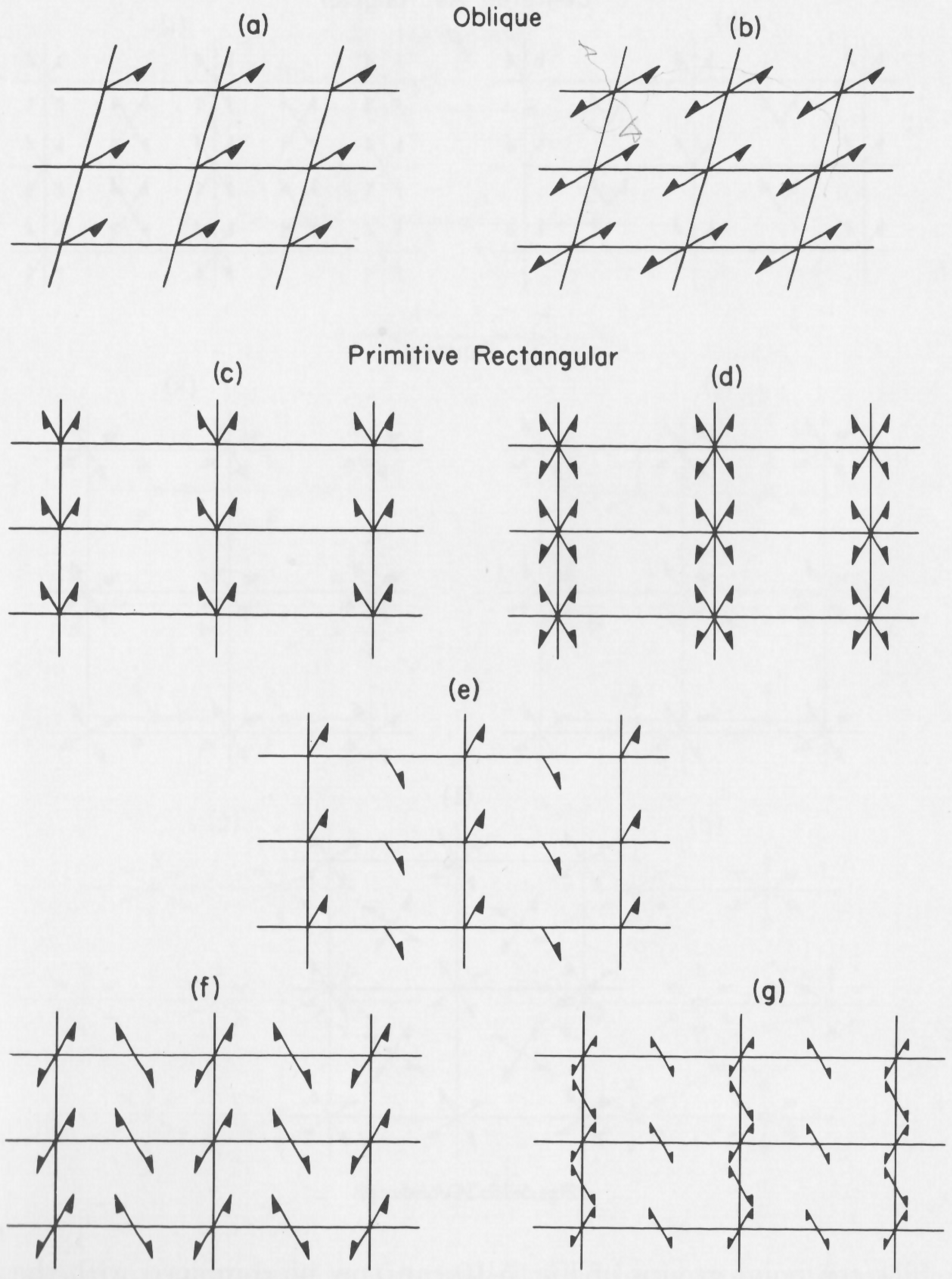


FIG. 5-12. The seventeen plane lattices.

It is also possible in a plane lattice to have a symmetry operation, called a *glide line,* which combines a translation and a reflection; this is illustrated in Fig. 5-11. Note that the glide operation introduces the reflected arrow at points equal to half a lattice constant. This is necessary so that the second application of the glide returns the arrow to a position equivalent to its original one.

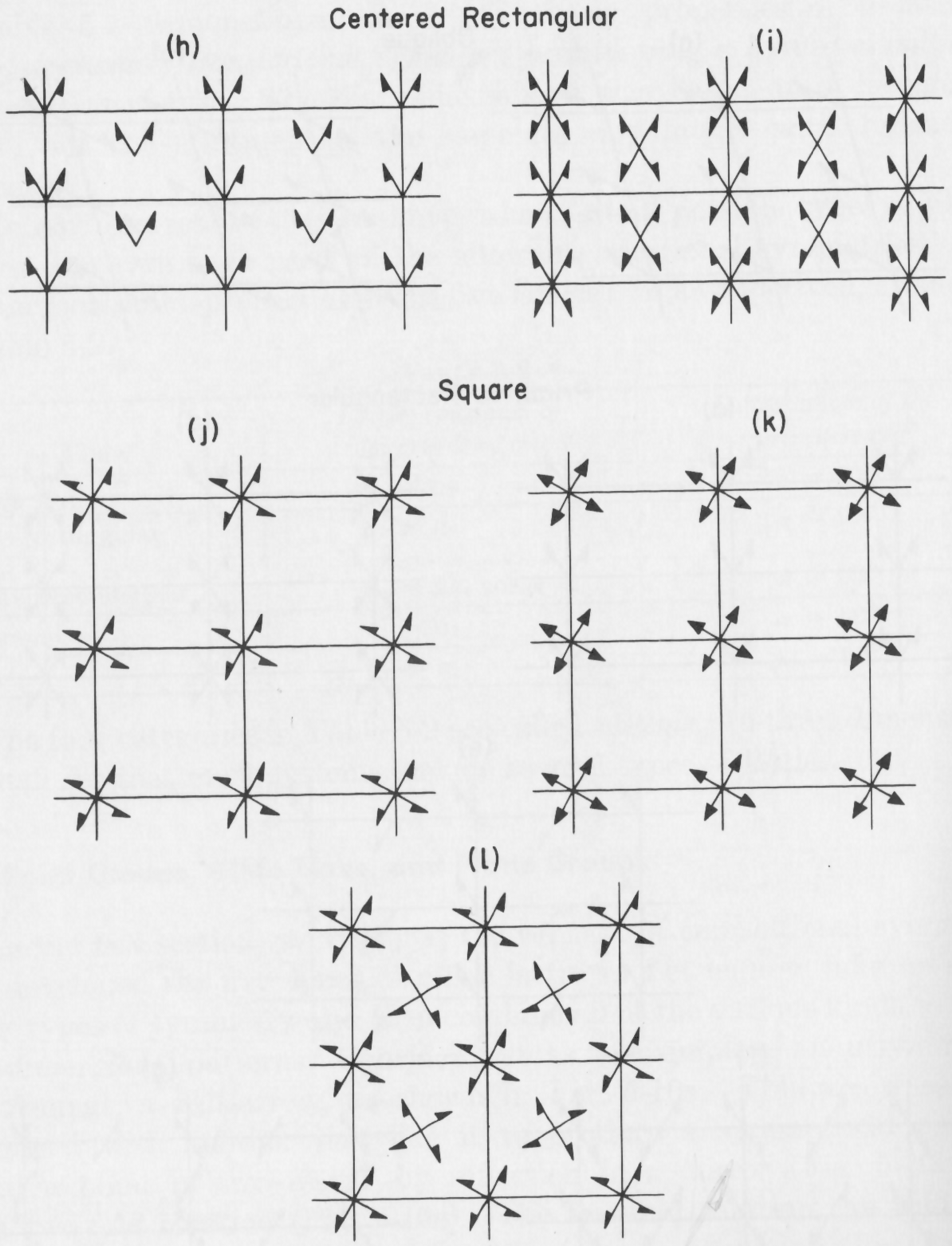


Fig. 5-12 (Continued).

The ten point groups of Fig. 5-10 can now be combined with the five plane lattices to form patterns called *plane groups*. Although 5 × 10 = 50, there are, in fact, only 13 possible plane groups obtained this way, because of the requirement that the arrangement placed at each lattice point be *consistent* with the lattice. This is a difficult term to define, so we shall show what it means by developing the plane groups. Let us start with the single unsymmetrical element of Fig. 5-10a. This element can be placed at each point of the oblique lattice (Fig. 5-3) to get the pattern of Fig. 5-12a. The two-fold point group (Fig. 5-10c), has the same symmetry as the

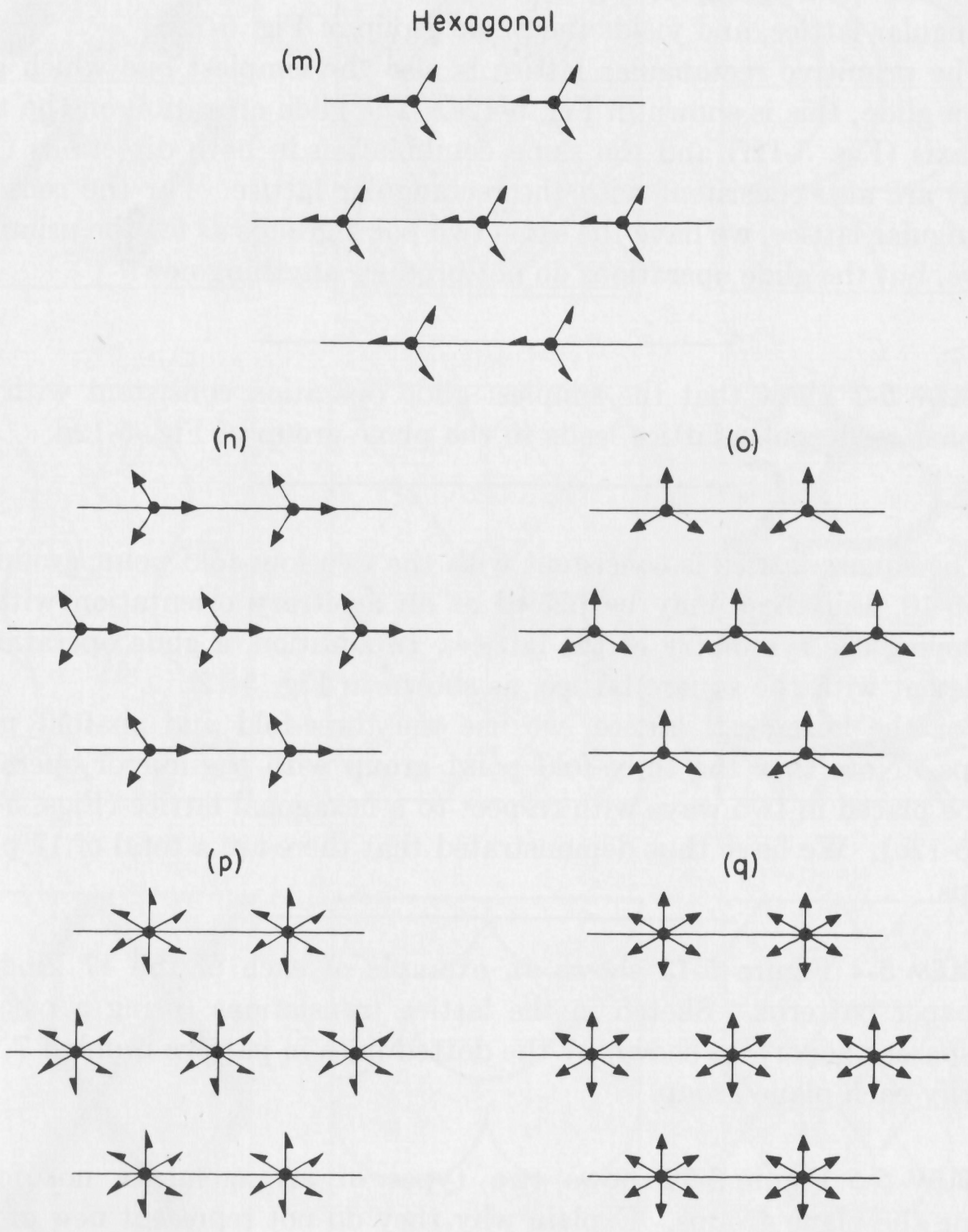


Fig. 5-12 (Continued).

inherent symmetry of the lattice, and it also leads to an oblique plane group (Fig. 5-12b). These are the only point groups consistent with the oblique lattice. A four-fold group, for example, could be placed at each lattice point, but the highest symmetry of the lattice is two-fold and it destroys the symmetry of the point group.

The point group of Fig. 5-10b involves only a mirror operation, and the lattice of lowest symmetry consistent with this is the rectangular one of Fig. 5-5. Note that the mirror line must be parallel to one of the two primitive vectors, so the resultant arrangement appears as shown in Fig.

5-12c. The point group of Fig. 5-10d is also consistent with the primitive rectangular lattice, and yields the plane group of Fig. 5-12d.

The primitive rectangular lattice is also the simplest one which goes with a glide; this is shown in Fig. 5-12e. The glide operation on the two-fold axis (Fig. 5-12f) and the same combination in both directions (Fig. 5-12g) are also consistent with the rectangular lattice. For the centered rectangular lattice, we have the same two point groups as for the primitive lattice, but the glide operations do not produce anything new.

PROBLEM 5-3 Show that the simplest glide operation consistent with the centered rectangular lattice leads to the plane group of Fig. 5-12d.

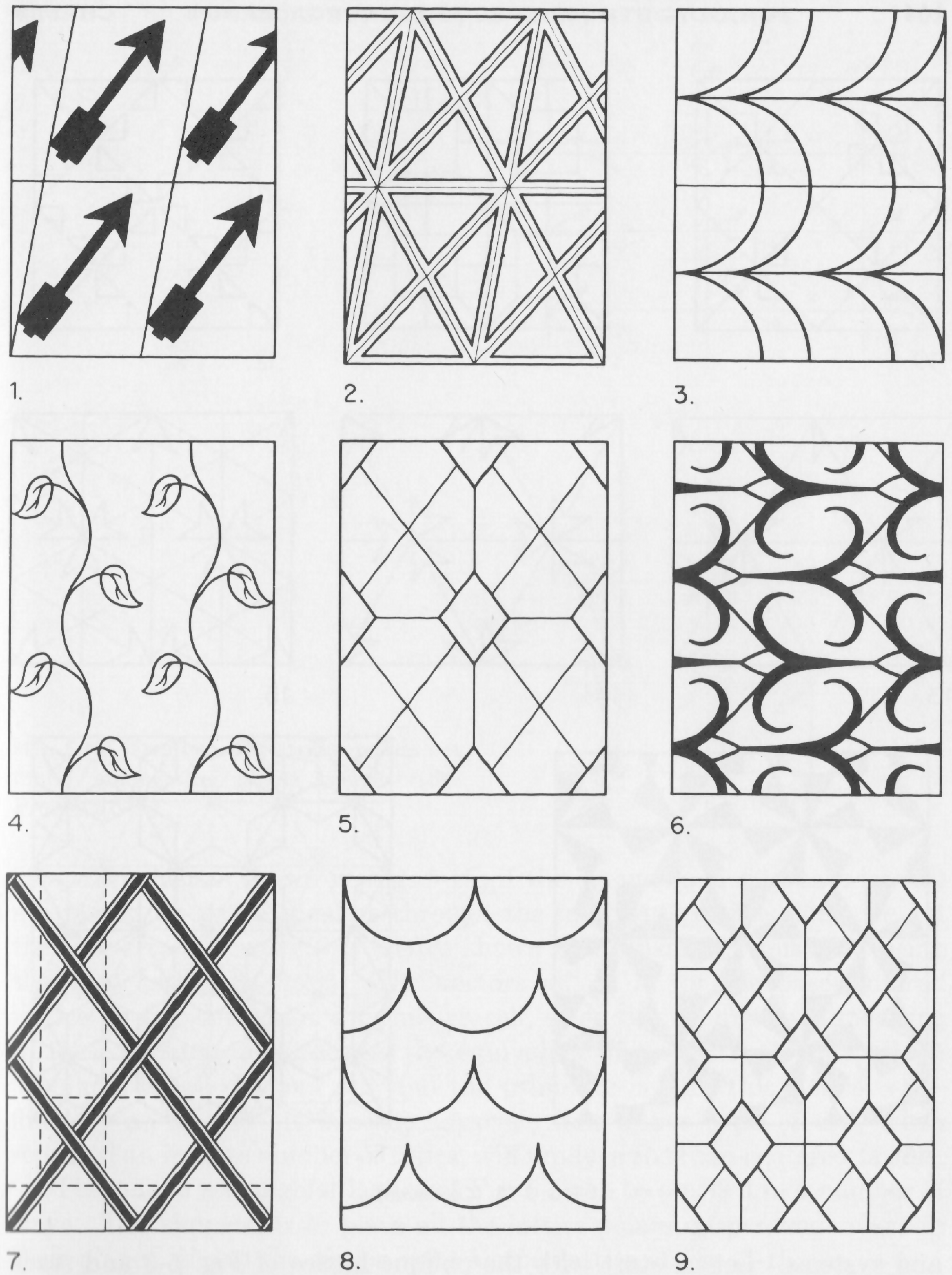
The square lattice is consistent with the two four-fold point groups of Fig. 5-10, and these may be placed at an arbitrary orientation without destroying the symmetry of the lattice. In addition, a glide operation is consistent with the square lattice, as shown in Fig. 5-12l.

For the hexagonal lattice, we use the three-fold and six-fold point groups. Note that the three-fold point group with the mirror operation can be placed in two ways with respect to a hexagonal lattice (Figs. 5-12n and 5-12o). We have thus demonstrated that there are a total of 17 plane groups.

PROBLEM 5-4 Figure 5-13 shows an example of each of the 17 kinds of wallpaper patterns. Sketch in the lattice translations (using a piece of transparent paper), as shown by the dotted lines in pattern number 7, and identify each plane group.

PROBLEM 5-5 Figure 5-14 shows two types of arrangements not found among the plane groups. Explain why they do not represent new groups and identify the group or groups to which they belong.

One important item to point out in concluding our study of plane groups is the following: it should be noted in Fig. 5-12 that each point group is consistent with one, and only one, type of lattice, whether used alone or in conjunction with a glide operation. For example, seven rectangular groups are formed from two point groups, these being either the mirror or the mirror plus a two-fold rotation. This same principle will also be found to hold for three-dimensional lattices, which we shall discuss in the next section.

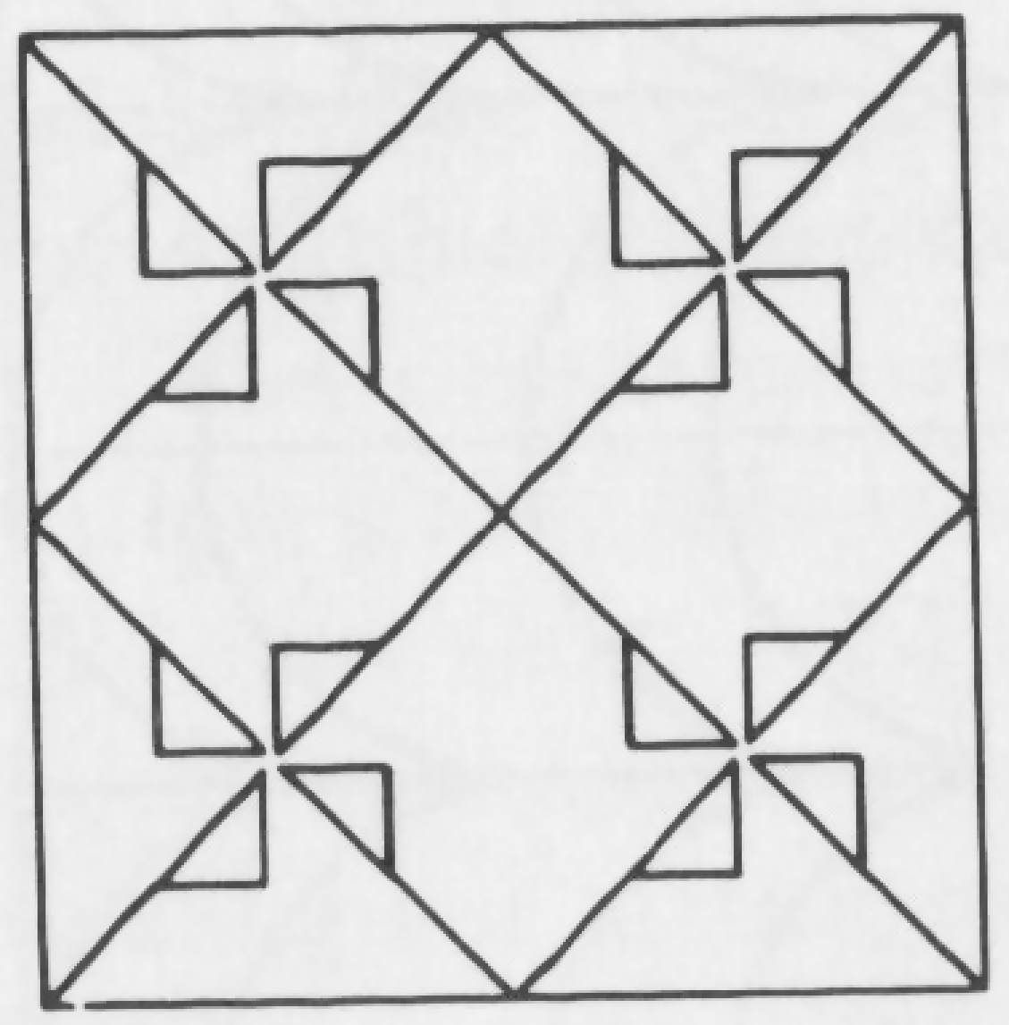


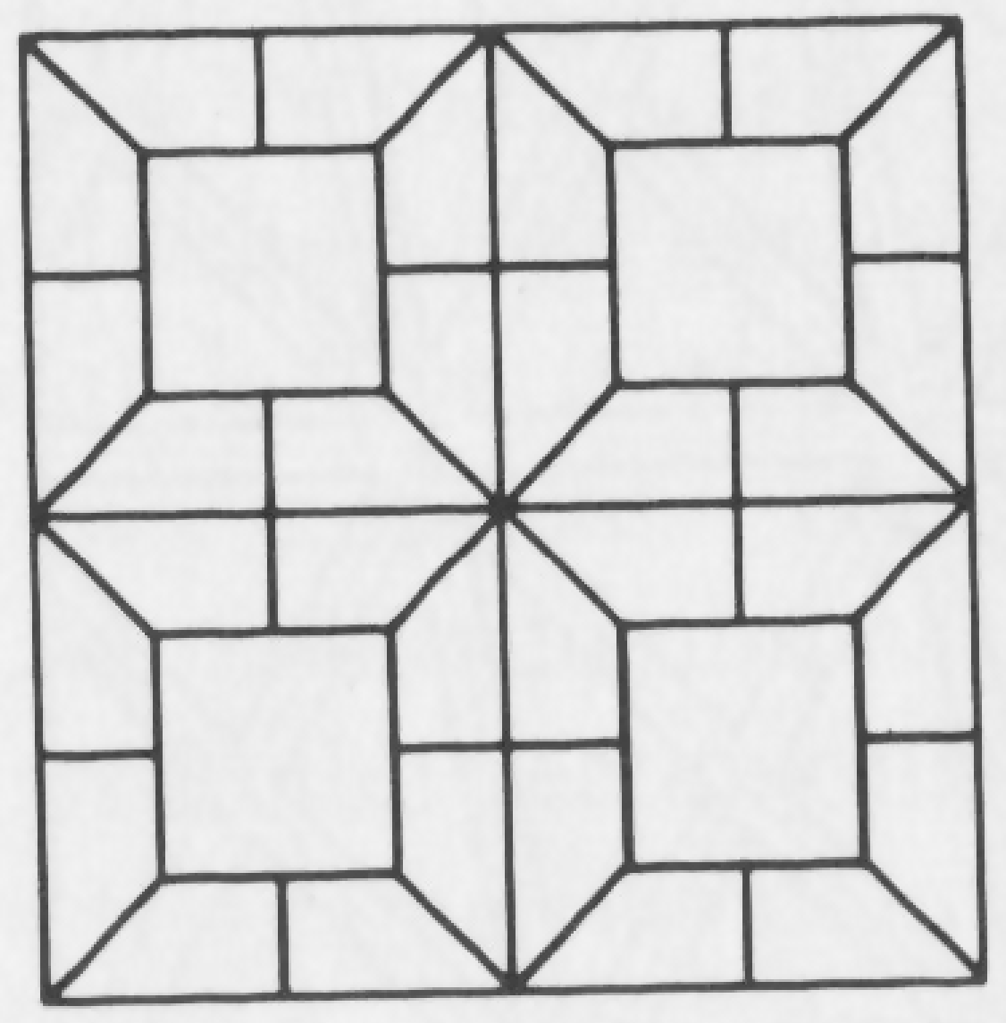
1. 2. 3. 4. 5. 6. 7. 8. 9.

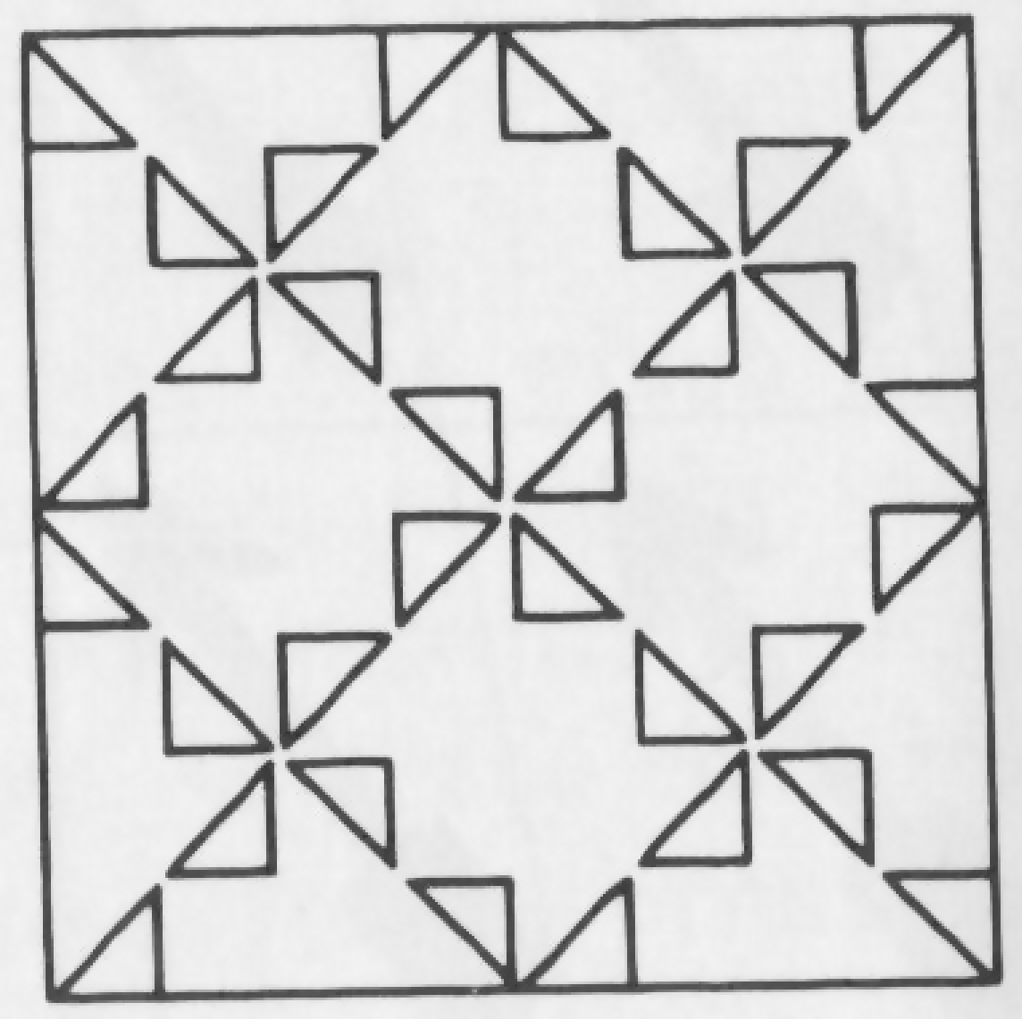
FIG. 5-13. Wallpaper patterns as examples of plane lattices. (From Journ. Text. Inst., with permission.)

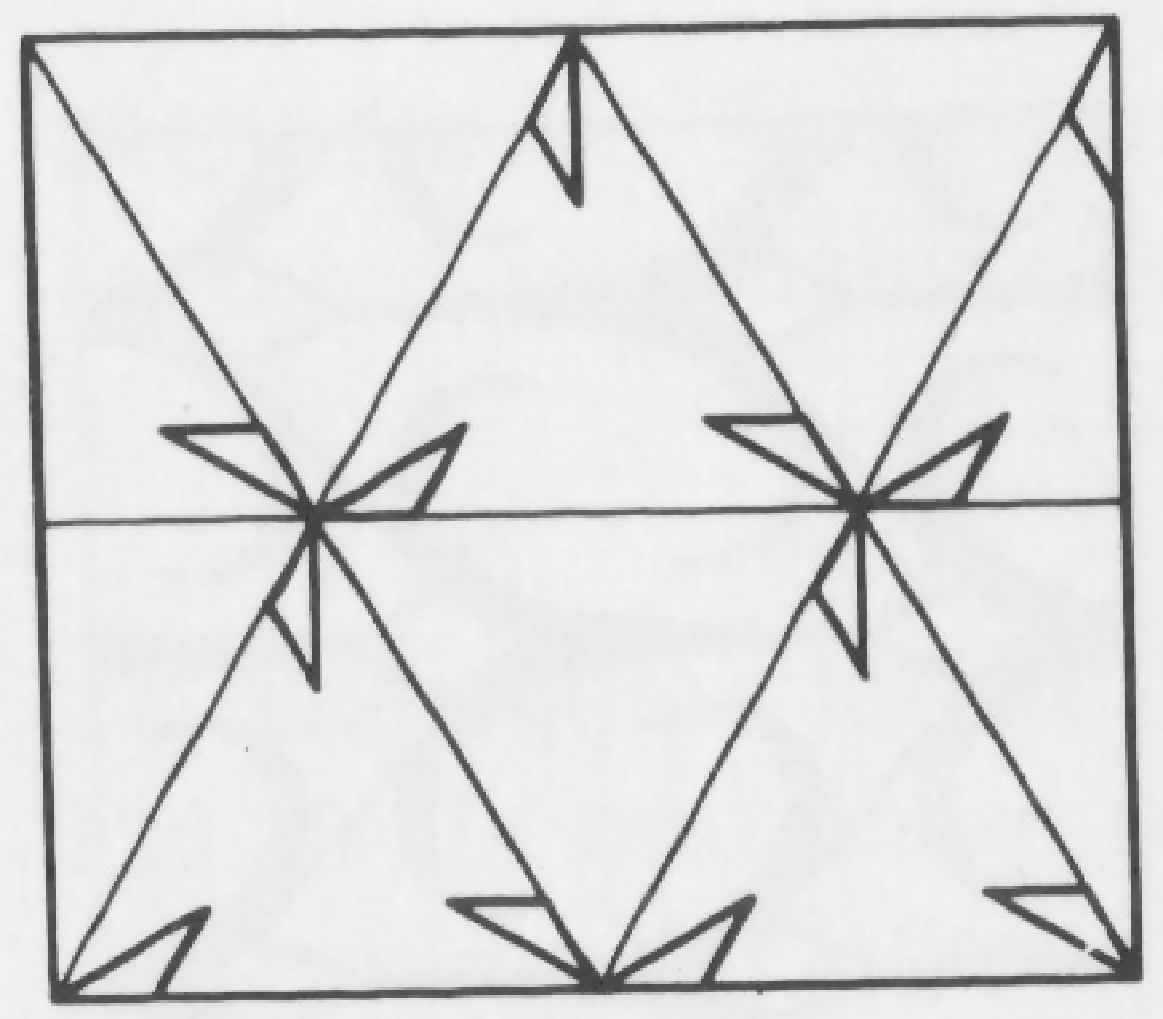
5-3 The Bravais Lattices

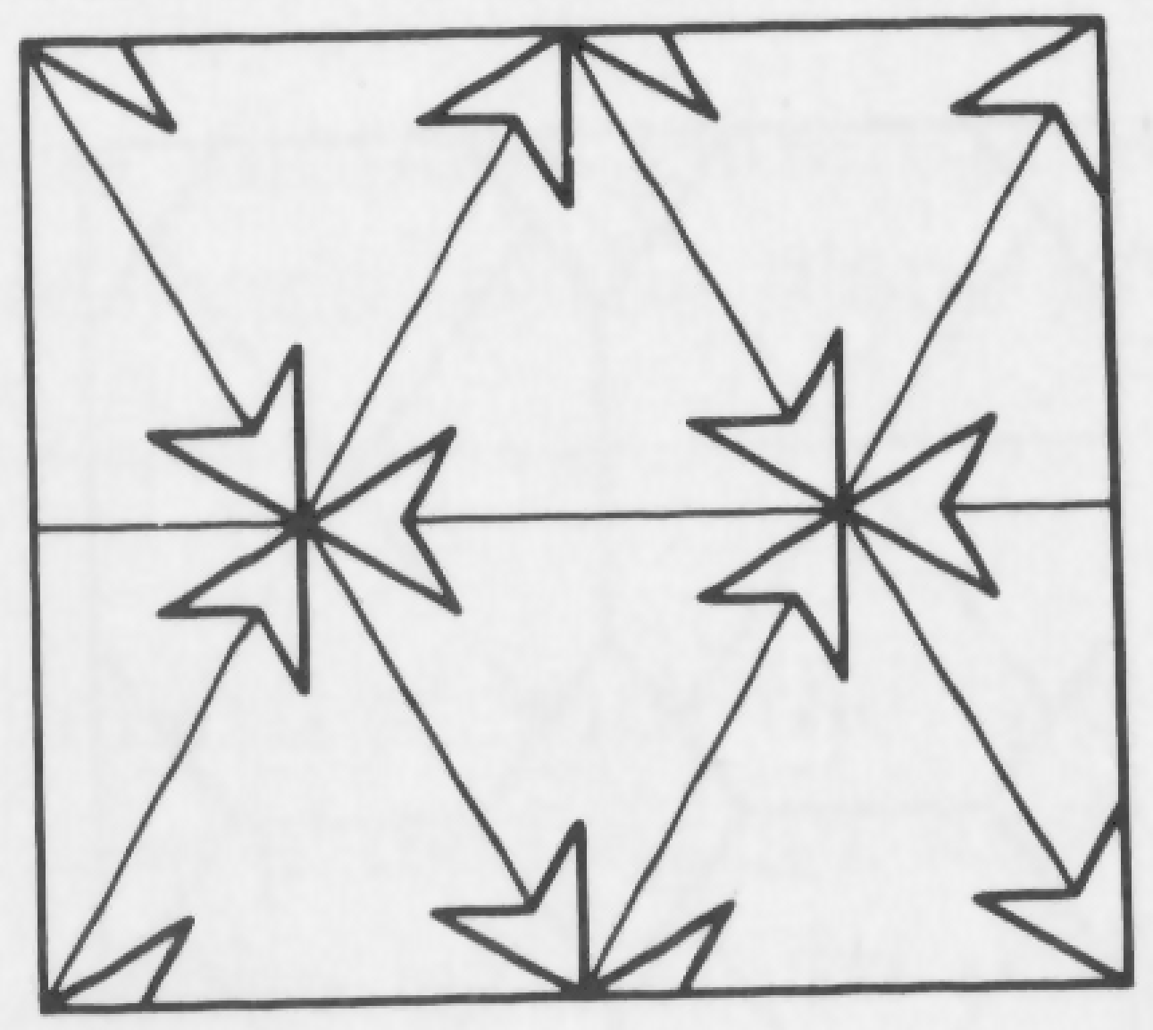
The problem of enumerating the possible three-dimensional arrangements is similar to the two-dimensional situation we have just covered. We shall reverse the order, however, and first discuss the possible lattices

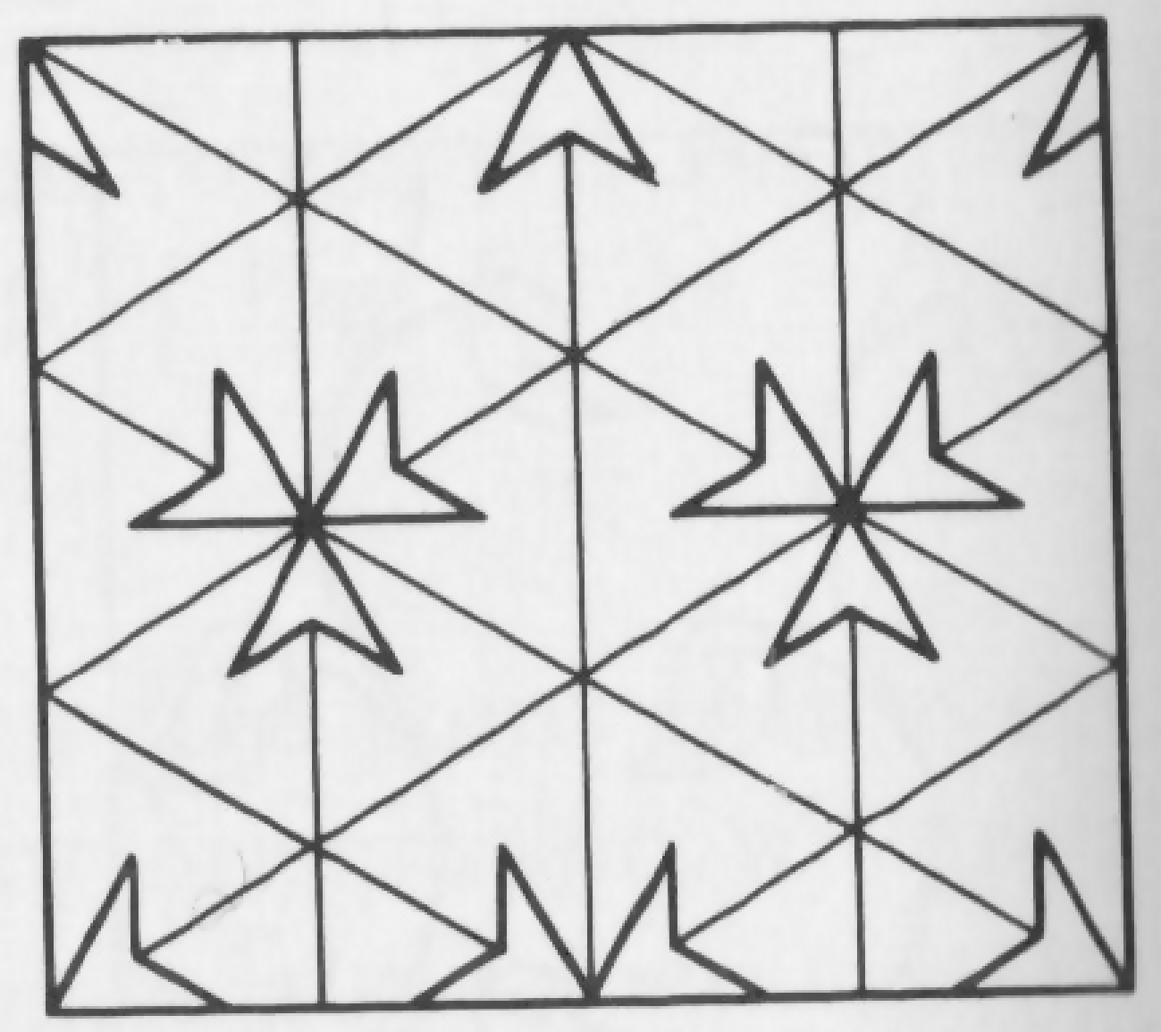

10.

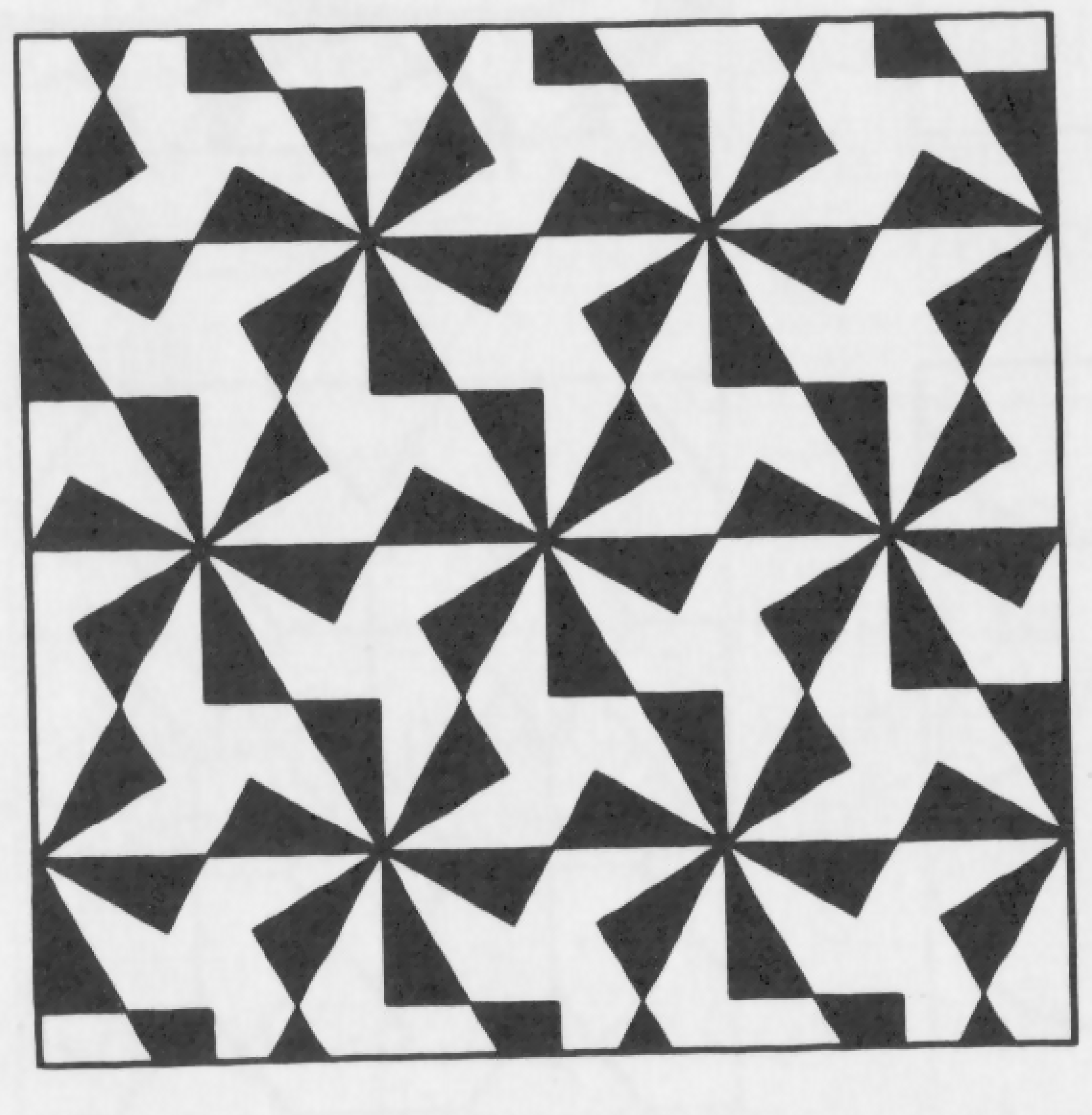

11.


12.


13


14.


15.


16.

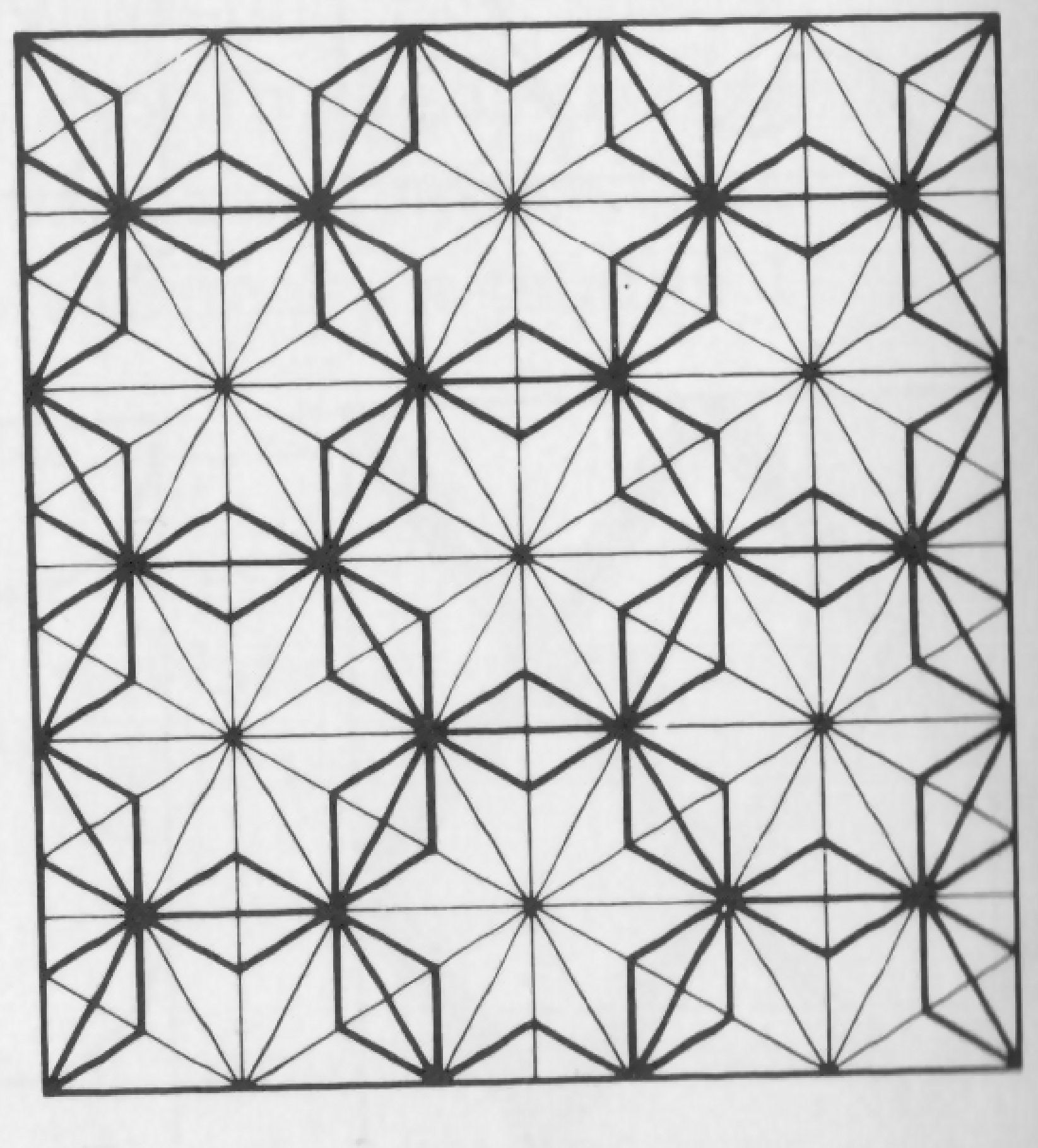

17.

FIG. 5-13 (Continued).

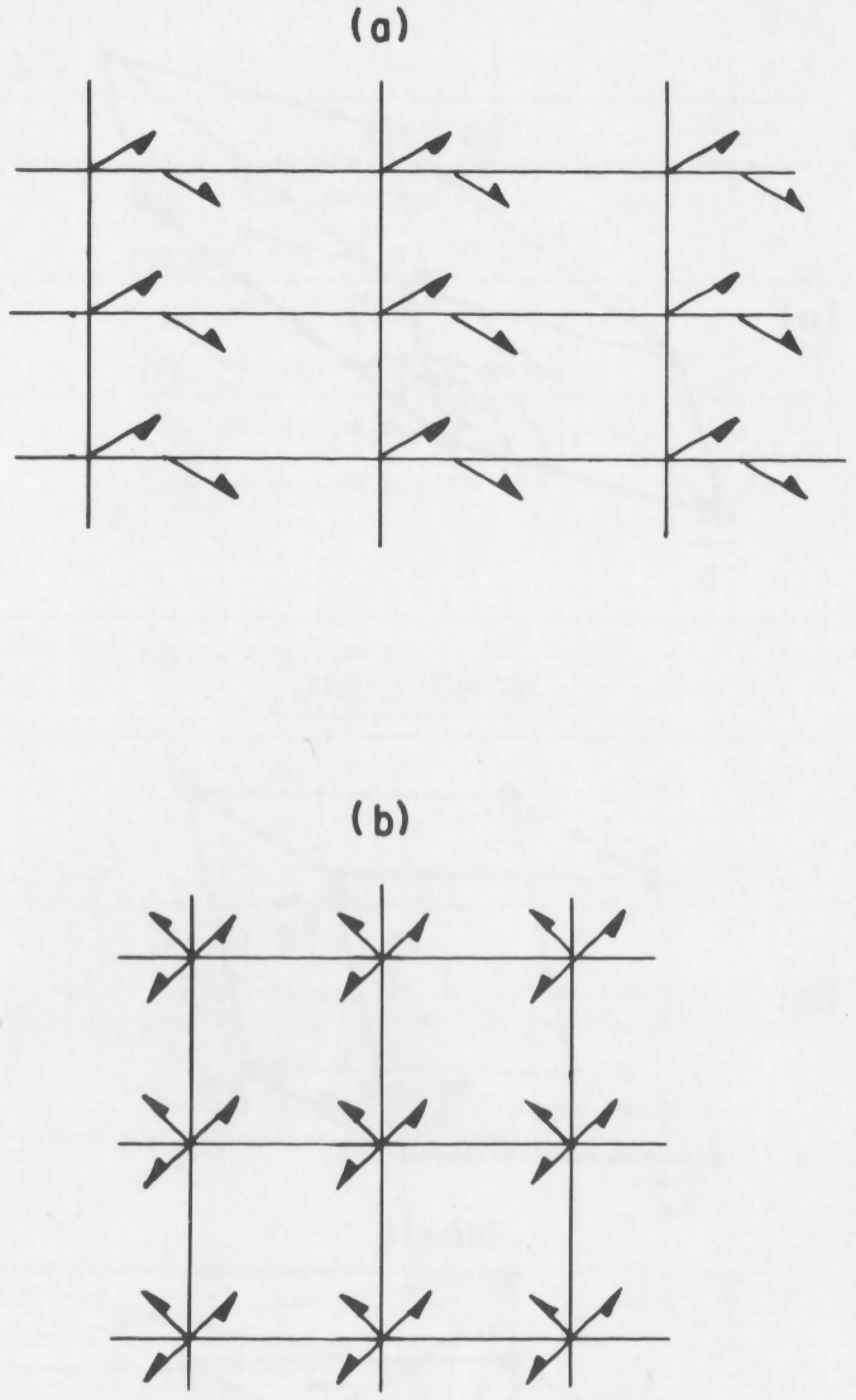


FIG. 5-14. Two types of arrangements not found among the plane groups. (See Problem 5-5.)

and systems. Let us start with the oblique lattice of Fig. 5-3 and stack similar lattices in the most general fashion possible; that is, in Fig. 5-15, we build up the three-dimensional network under the conditions

$$a_1 \neq a_2 \neq a_3 \qquad \alpha_1 \neq \alpha_2 \neq \alpha_3 \neq 90 \text{ deg}$$

This arrangement is called a *triclinic* lattice. If the oblique lattices are stacked so that the points of the second and succeeding layers lie on perpendiculars through the points of the original layer, we get the simple *monoclinic* lattice shown in Fig. 5-15; if the second (and alternate layers) are placed on perpendiculars through the midpoints of the sides, we get the *base-centered monoclinic* lattice shown (the "base" in this case being the front and back faces). The vectors $\mathbf{a}_1$, $\mathbf{a}_2$, $\mathbf{a}_3$ for the base-centered lattice do not determine a primitive cell, since the volume corresponding to these three vectors encloses the equivalent of *two* lattice points. Such a volume is called a *unit cell,* and the primitive cell for this lattice is defined by $\mathbf{a}_1$, $\mathbf{a}_2$, $\mathbf{a}_3'$. It is seen, however, that either type of cell, when repeated an infinite number of times, will produce the base-centered lattice.

The simple rectangular lattice of Fig. 5-5 can be stacked in a number of ways. The simplest is to place all the lattice points on perpendiculars to their planes, giving the *simple orthorhombic* lattice of Fig. 5-15.

It may be asked at this point why the angles α_1 and α_2 cannot be arbitrary instead of 90 deg; the answer is that the resulting lattice would not be consistent with any of the 32 possible point groups which, we shall soon show, exist in three dimensions. The second possible way of stacking the rectangular lattice gives a base-centered arrangement of the same type as that which we obtained above. However, if the lattice points of the second net fall over the centers of the lattices, we obtain the *body-centered orthorhombic* lattice.

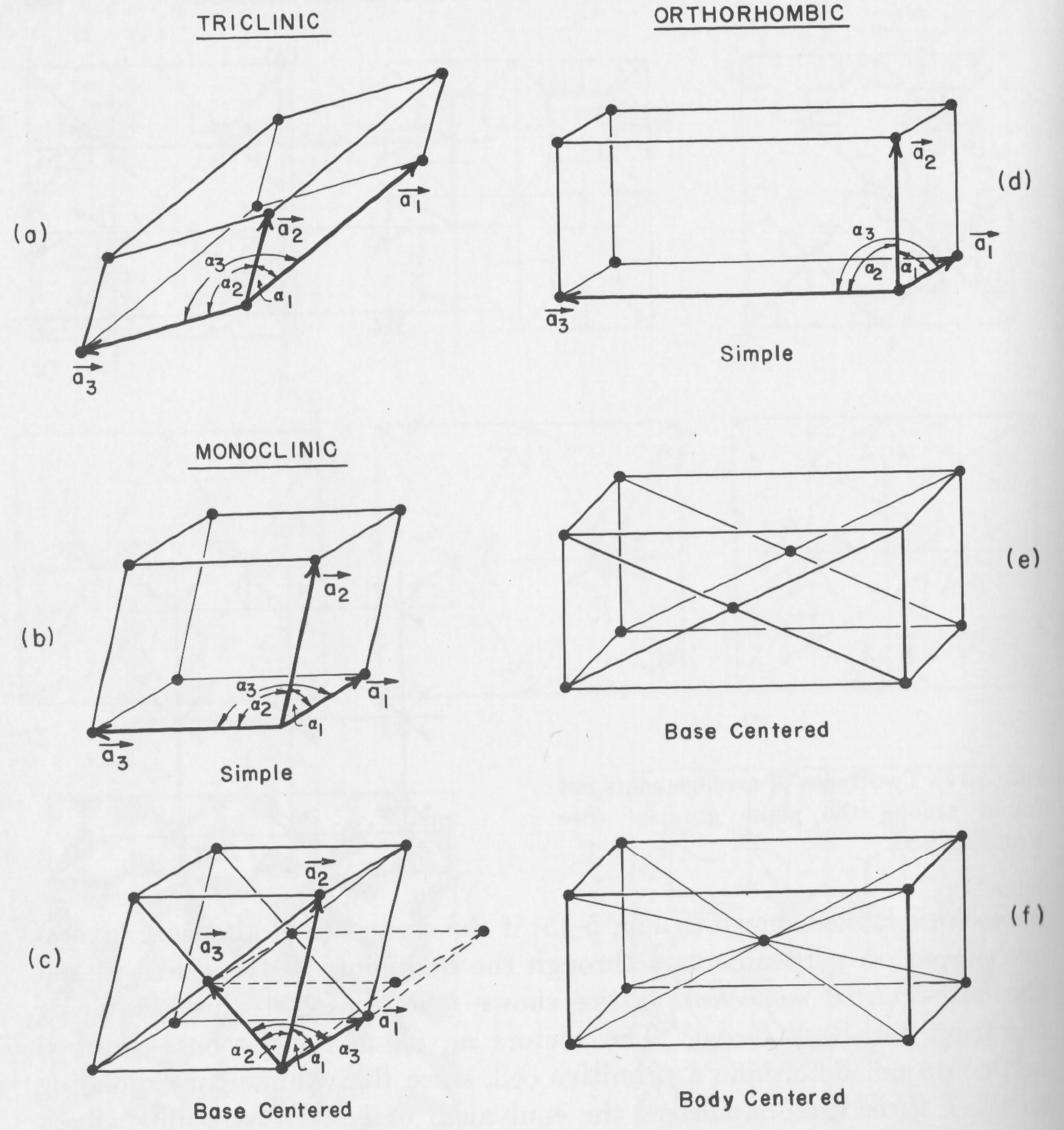

TRICLINIC
ORTHORHOMBIC
(a)
$\vec{a}_1$
$\vec{a}_2$
$\vec{a}_3$
α_1
α_2
α_3
(d)
Simple
MONOCLINIC
(b)
(e)
Simple
Base Centered
(c)
(f)
Base Centered
Body Centered

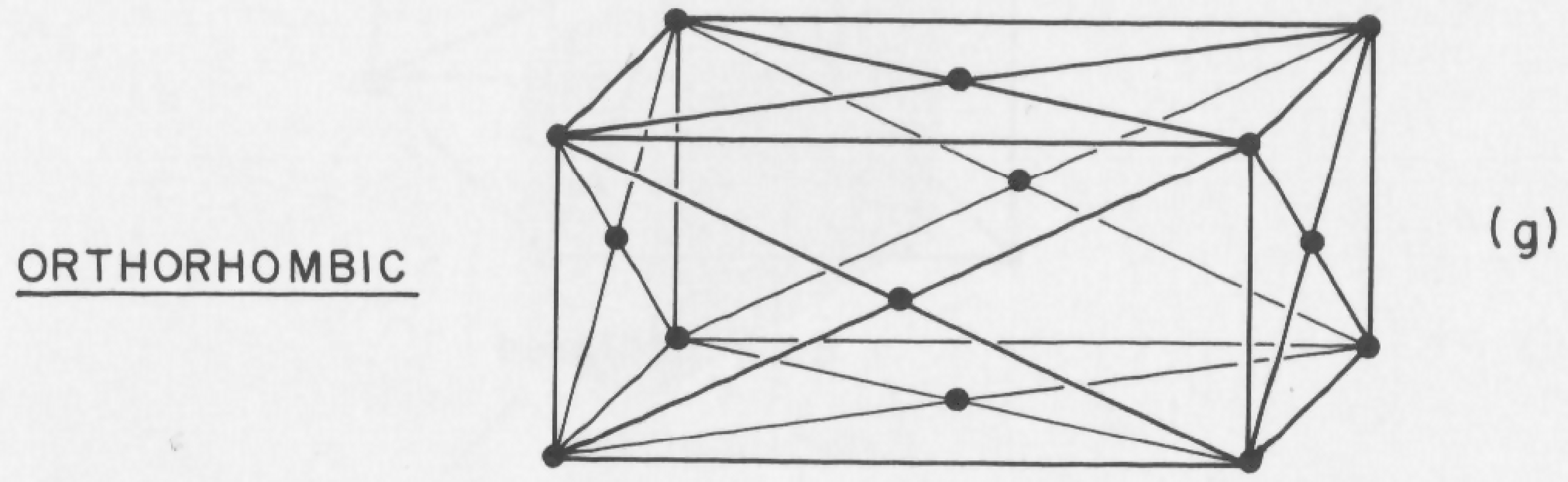
FIG. 5-15. The Bravais lattices.

Square lattices can be stacked to form a *simple tetragonal* or a *body-centered tetragonal* lattice, if $a_1 = a_2 \neq a_3$, and to form a *simple, body-centered,* or *face-centered cubic* lattice when $a_1 = a_2 = a_3$. Finally, the hexagonal nets can be stacked in two ways. If three layers are arranged as shown in Fig. 5-15m (the fourth layer being directly over the first layer), the total symmetry is reduced to three-fold and results in a *trigonal* lattice. If the lattice points are one above the other, we have a *hexagonal* lattice.

To summarize, then, we have demonstrated the existence of 14 translational arrangements, called *Bravais lattices;* these lattices can be divided into seven systems. Table 5-3 lists the properties of the unit cell of each of the systems.

TABLE 5-3

System	*Bravais lattice*	*Unit cell*
1. Triclinic	1. Simple	$a_1 \neq a_2 \neq a_3$ $\alpha_1 \neq \alpha_2 \neq \alpha_3 \neq 90°$
2. Monoclinic	2. Simple	$a_1 \neq a_2 \neq a_3$
	3. Base-centered	$\alpha_1 = \alpha_2 = 90° \neq \alpha_3$
3. Orthorhombic	4. Simple	$a_1 \neq a_2 \neq a_3$
	5. Base-centered	$\alpha_1 = \alpha_2 = \alpha_3 = 90°$
	6. Body-centered	
	7. Face-centered	
4. Tetragonal	8. Simple	$a_1 = a_2 \neq a_3$
	9. Body-centered	$\alpha_1 = \alpha_2 = \alpha_3 = 90°$
5. Cubic	10. Simple	$a_1 = a_2 = a_3$
	11. Face-centered	$\alpha_1 = \alpha_2 = \alpha_3 = 90°$
	12. Body-centered	
6. Trigonal	13. Simple	$a_1 = a_2 = a_3$ $120° > \alpha_1 = \alpha_2 = \alpha_3 \neq 90°$
7. Hexagonal	14. Simple	$a_1 = a_2 \neq a_3$ $\alpha_1 = 120°, \quad \alpha_2 = \alpha_3 = 90°$


ORTHORHOMBIC
(g)

Face Centered

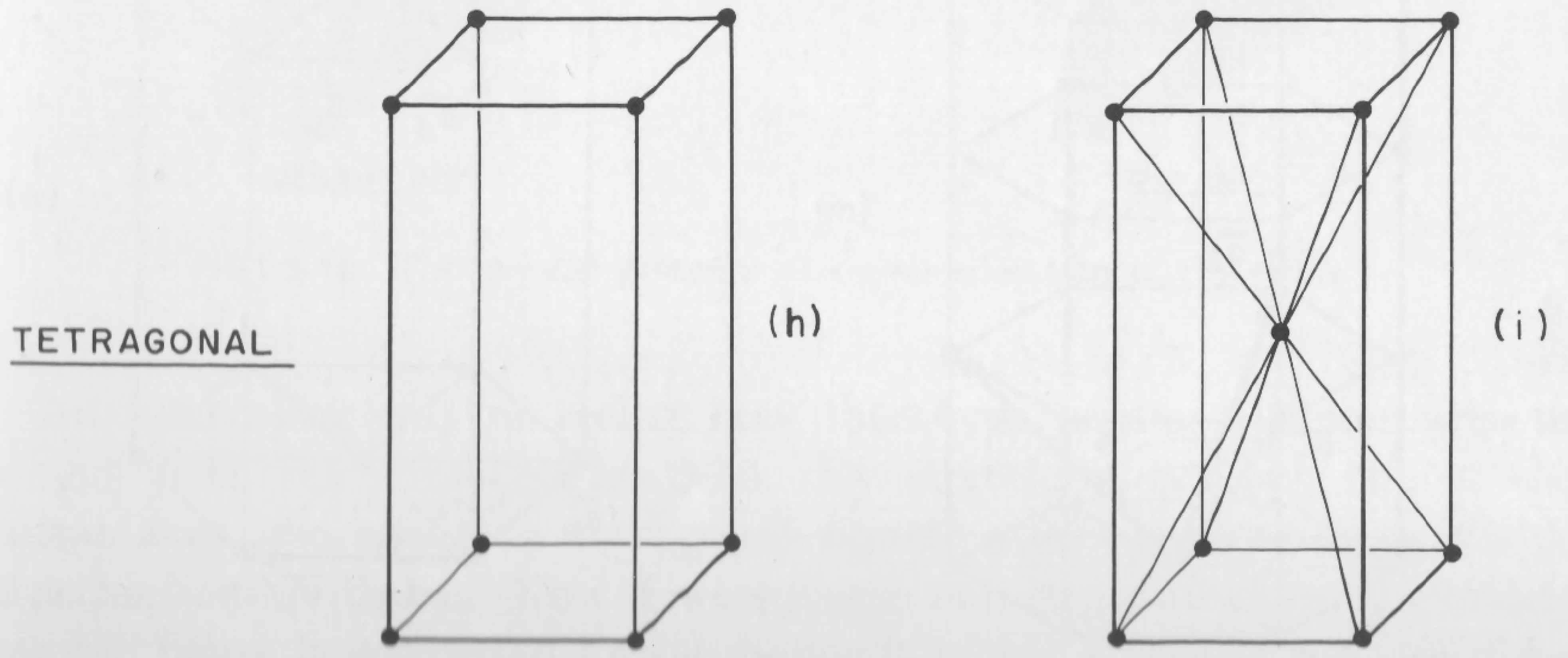

TETRAGONAL
(h)
(i)

Fig. 5-15 (Continued). Simple Body Centered

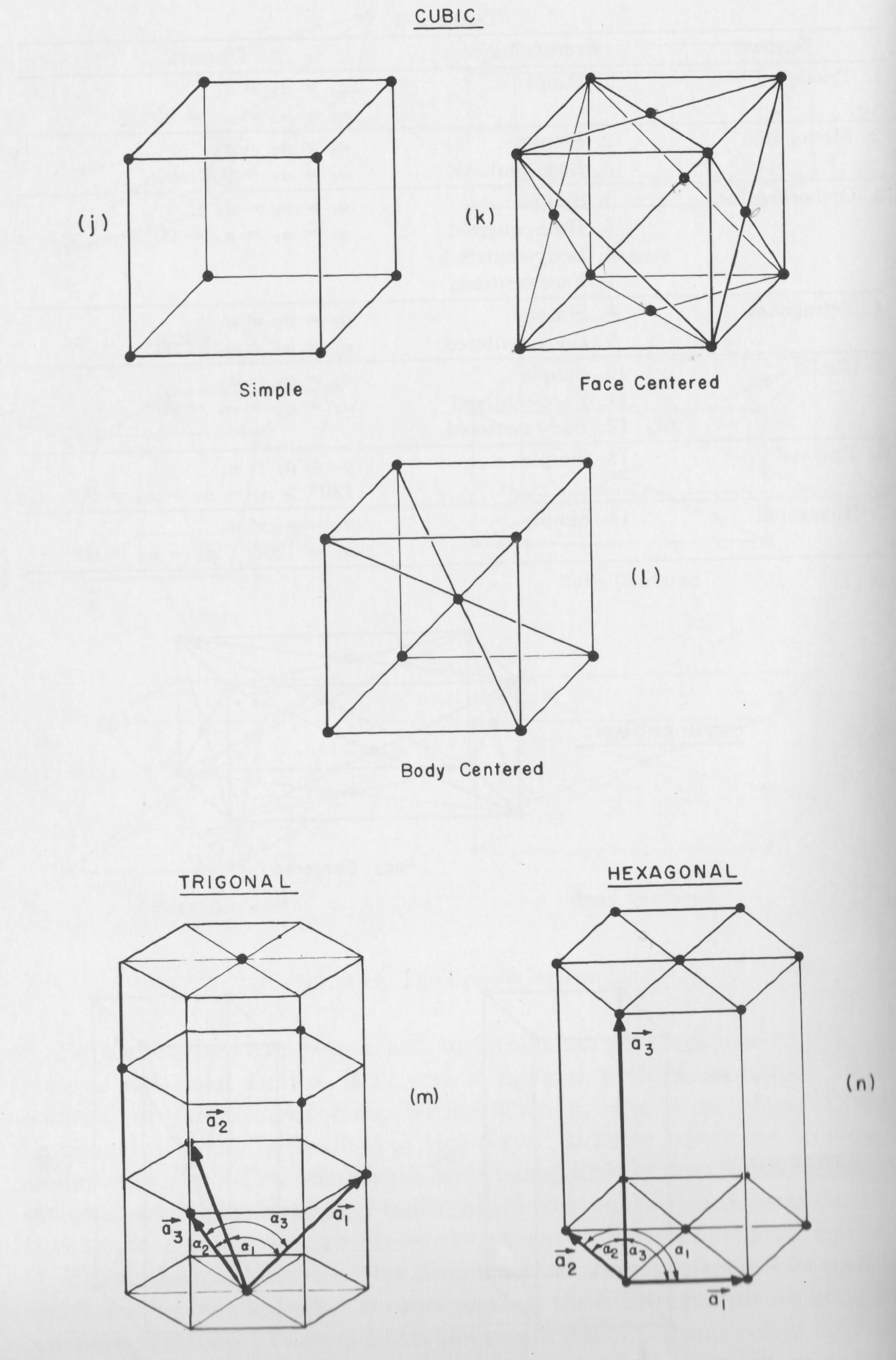

CUBIC
(j)
Simple
(k)
Face Centered
(l)
Body Centered
TRIGONAL
(m)
$\vec{a}_2$
$\vec{a}_3$
$\vec{a}_1$
α_3
α_2
α_1
HEXAGONAL
(n)
$\vec{a}_3$
$\vec{a}_2$
$\vec{a}_1$
α_2
α_3
α_1

Fig. 5-15 (Continued).

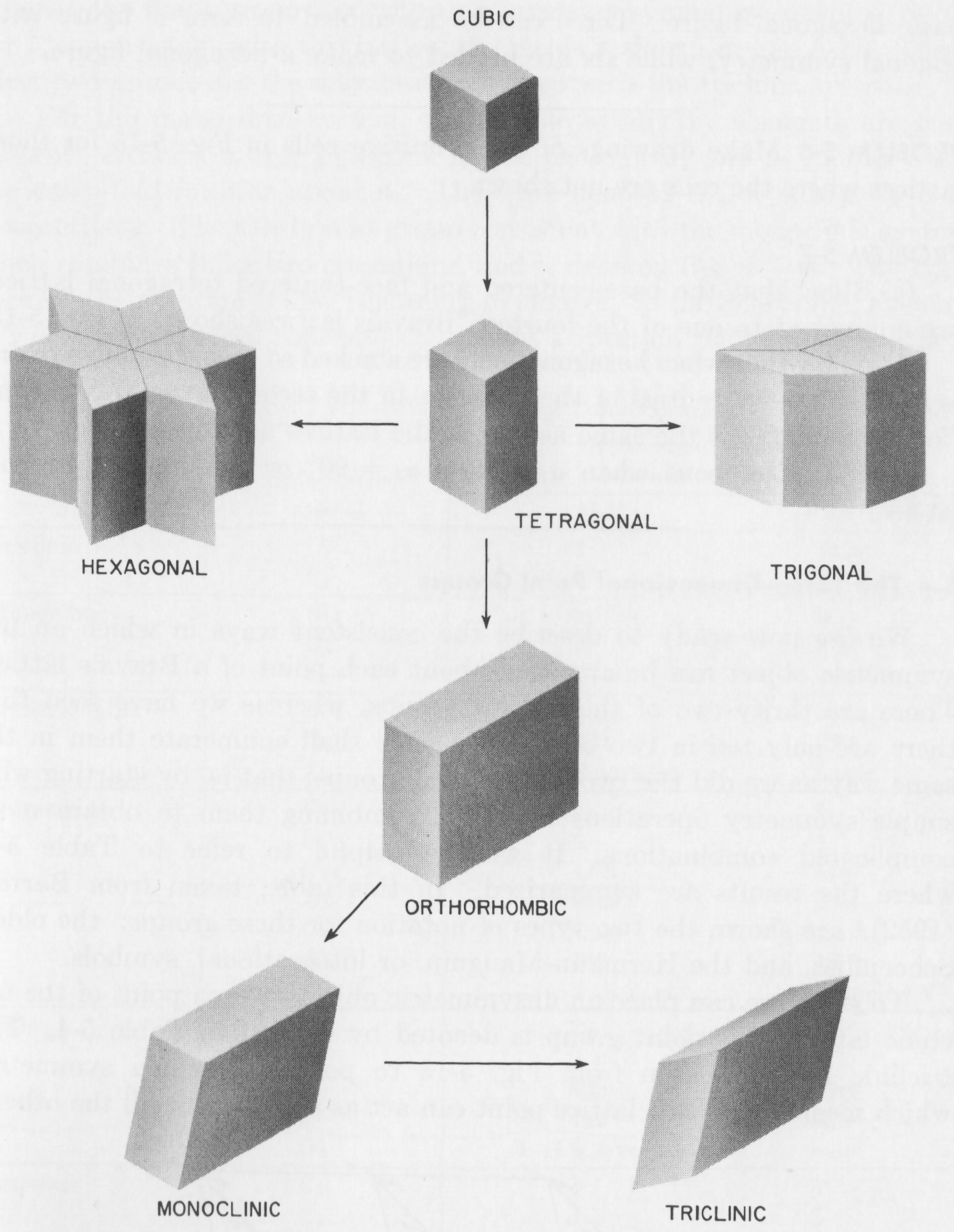

CUBIC
HEXAGONAL
TETRAGONAL
TRIGONAL
ORTHORHOMBIC
MONOCLINIC
TRICLINIC

FIG. 5-16. The crystal systems as a generalization of the cube.

An interesting way of seeing how the seven crystal systems arise is to start with the cube of Fig. 5-16. By stretching it along one of the central axes, we obtain a tetragonal figure; a second stretching yields an orthorhombic figure. Now if we squeeze two opposite sides, the orthorhombic figure is converted into a monoclinic one; a second squeeze gives a triclinic shape. If we return to the tetragonal figure, a squeeze gives the basic hexagonal figure. Three can be assembled to form a figure with trigonal symmetry, while six are needed to make a hexagonal figure.

PROBLEM 5-6 Make drawings of the primitive cells in Fig. 5-15 for those lattices where the cells are not shown.

PROBLEM 5-7

(a) Show that the base-centered and face-centered tetragonal lattices are equivalent to one of the fourteen Bravais lattices shown in Fig. 5-15.

(b) Show that when hexagonal nets are stacked so that the points of one layer lie over the centers of the triangles in the second layer, the translational symmetry is the same as one of the lattices already defined.

(c) What happens whan $\alpha_1 = \alpha_2 = \alpha_3 = 90°$ or $120°$ in the trigonal lattice?

5-4 The Three-Dimensional Point Groups

We are now ready to describe the consistent ways in which an unsymmetric object can be arranged about each point of a Bravais lattice. There are thrity-two of these point groups, whereas we have seen that there are only ten in two dimensions. We shall enumerate them in the same way as we did the two-dimensional groups; that is, by starting with simple symmetry operations and then combining them to obtain more complicated combinations. It will be helpful to refer to Table 5-4, where the results are summarized. In this table, taken from Barrett (1952),[2] are shown the two types of notation for these groups: the older, Schoenflies, and the Hermann-Mauguin, or international, symbols.

To start, we can place an unsymmetric object at each point of the triclinic lattice; this point group is denoted by C_1 or 1 in Table 5-4. The triclinic lattice is seen from Fig. 5-15 to possess inversion symmetry, which means that each lattice point can act as a mirror for all the others.

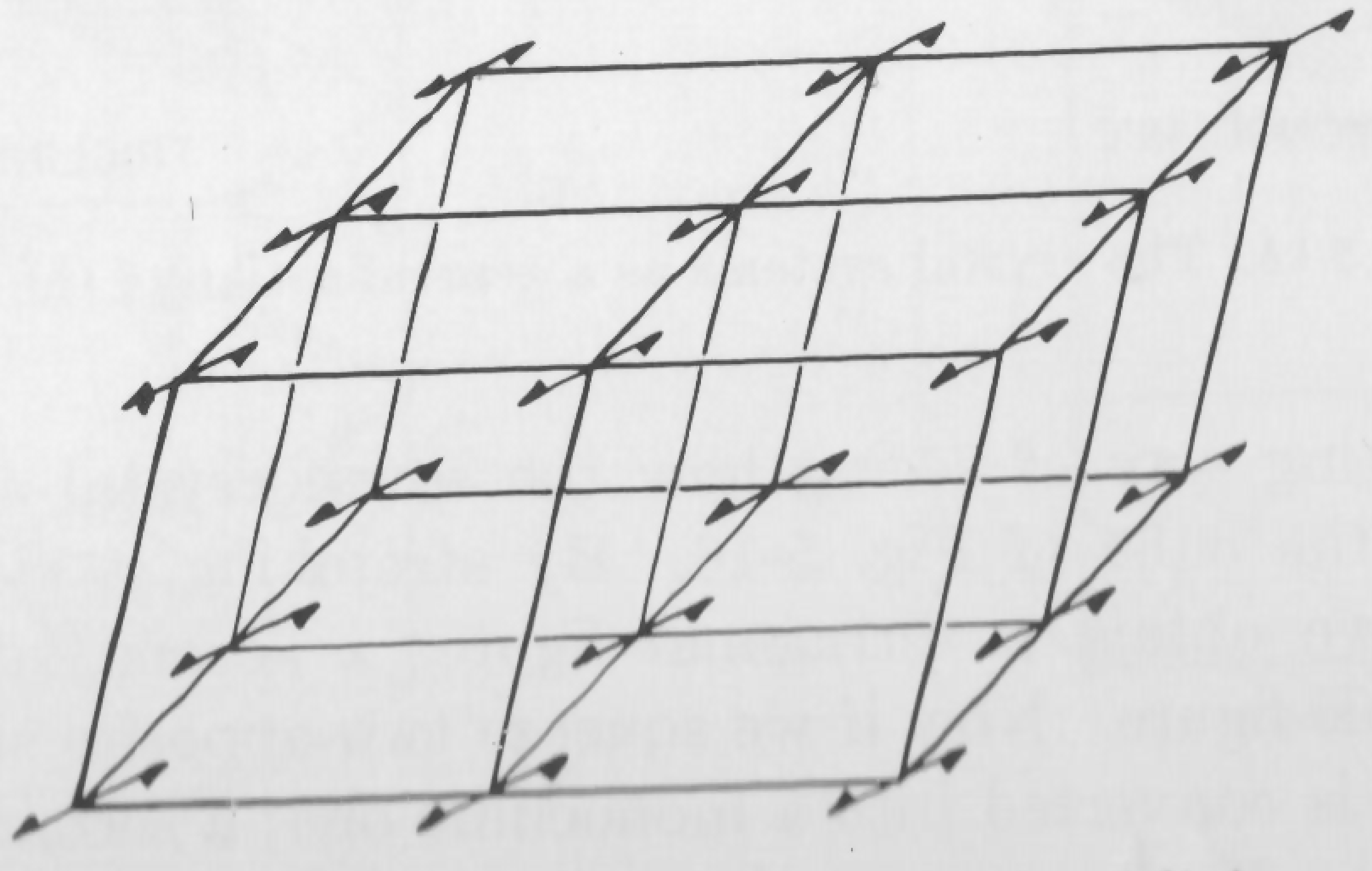


FIG. 5-17. A space group consistent with the triclinic system.

Hence, the point group incorporating inversion symmetry, denoted by C_i or $\bar{1}$, is consistent with this lattice; the result is shown in Fig. 5-17. These first two groups are the only ones which go with the triclinic system.

For the monoclinic system, two simple symmetry elements are consistent: either a mirror reflection in the plane of $\mathbf{a}_1$ and $\mathbf{a}_2$ in Fig. 5-15, or a two-fold rotation about $\mathbf{a}_3$. These are denoted C_{1h} or m and C_2 or 2, respectively. The third point group consistent with the monoclinic system then combines these two operations, and is denoted C_{2h} or $2/m$. The logic behind the notation is now beginning to emerge: in the Schoenflies system, C_n denotes an n-fold rotation, C_{nh} means a rotation plus a reflection in a

TABLE 5-4. The 32 Crystal Point Groups

System	*Schoenflies symbol*	*Hermann-Mauguin* or *international symbol*	
		Full	*Abbreviated*
Triclinic	C_1	1	1
	C_i	$\bar{1}$	$\bar{1}$
Monoclinic	C_{1h}	m	m
	C_2	2	2
	C_{2h}	$2/m$	$2/m$
Orthorhombic	C_{2v}	$2mm$	mm
	D_2	222	222
	D_{2h}	$2/m\ 2/m\ 2/m$	mmm
Tetragonal	C_4	4	4
	S_4	$\bar{4}$	$\bar{4}$
	C_{4h}	$4/m$	$4/m$
	D_{2d}	$\bar{4}2m$	$\bar{4}2m$
	C_{4v}	$4mm$	$4mm$
	D_4	422	42
	D_{4h}	$4/m\ 2/m\ 2/m$	$4/mmm$
Cubic	T	23	23
	T_h	$2/m\ \bar{3}$	$m3$
	T_d	$\bar{4}3m$	$\bar{4}3m$
	O	432	432
	O_h	$4/m\ \bar{3}\ 2/m$	$m3m$
Trigonal	C_3	3	3
	C_{3i}	$\bar{3}$	$\bar{3}$
	C_{3v}	$3m$	$3m$
	D_3	32	32
	D_{3d}	$\bar{3}\ 2/m$	$\bar{3}m$
Hexagonal	C_6	6	6
	C_{3h}	$\bar{6}$	$\bar{6}$
	C_{6h}	$6/m$	$6/m$
	D_{3h}	$\bar{6}2m$	$\bar{6}2m$
	C_{6v}	$6mm$	$6mm$
	D_6	622	62
	D_{6h}	$6/m\ 2/m\ 2/m$	$6/mmm$

plane perpendicular to the axis; in the international notation, the corresponding symbols are n (where $n = 1, 2, 3, 4$, and 6) and n/m.

For the orthorhombic system, we can combine a two-fold axis with a reflection in a vertical plane; that is, a plane passing through the axis. This is denoted by C_{2v} or $2mm$.

PROBLEM 5-8 Show that this symmetry operation automatically provides for two mirror planes at right angles, thus making the notation $2mm$ rather than $2m$. Show why it is sufficient to use mm as the abbreviated notation in Table 5-4.

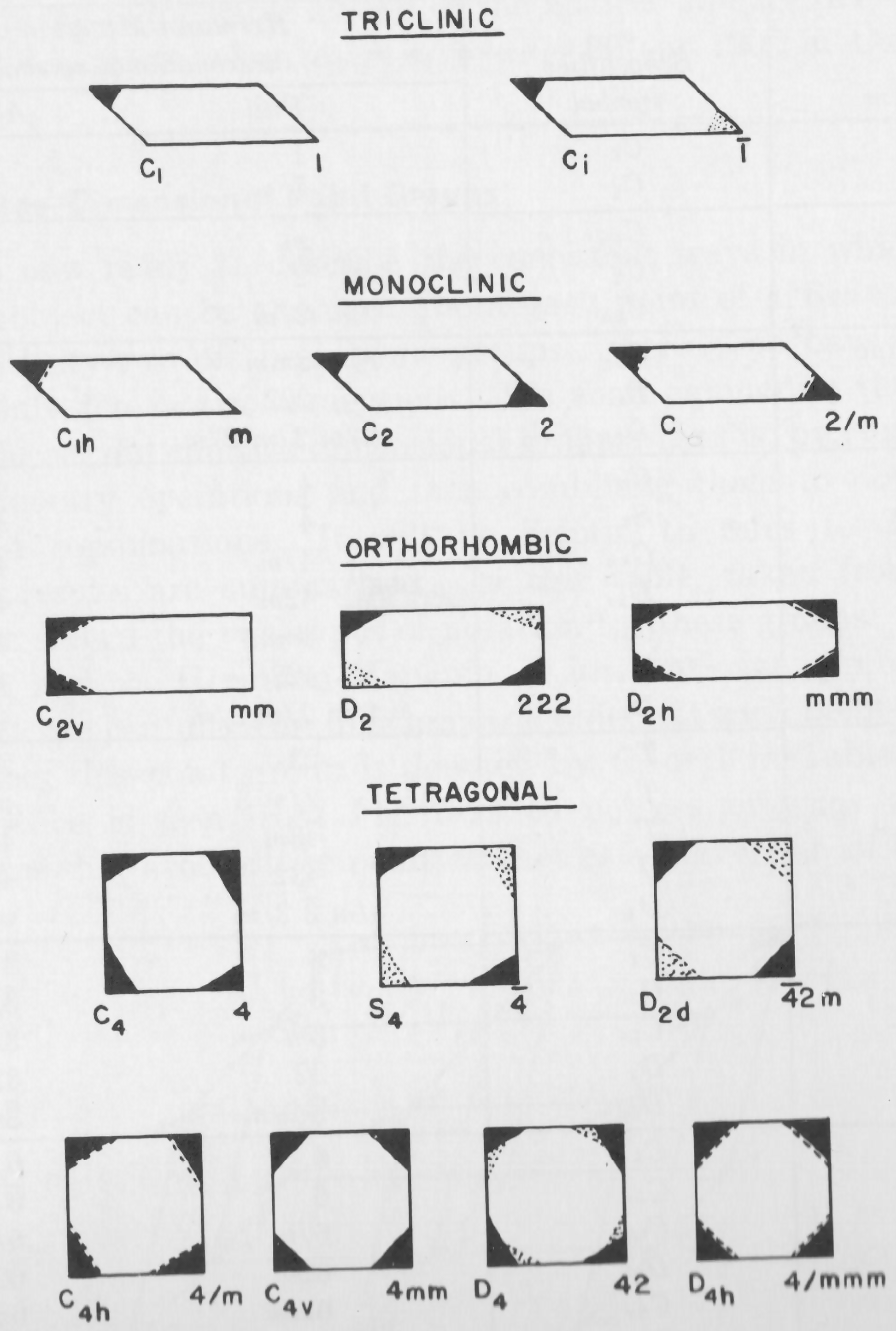

TRICLINIC
C_1 1
C_i $\bar{1}$
MONOCLINIC
C_{1h} m
C_2 2
C 2/m
ORTHORHOMBIC
C_{2v} mm
D_2 222
D_{2h} mmm
TETRAGONAL
C_4 4
S_4 $\bar{4}$
D_{2d} $\bar{4}2m$
C_{4h} 4/m
C_{4v} 4mm
D_4 42
D_{4h} 4/mmm

FIG. 5-18. The crystal point groups or classes for all systems except the cubic. (After L. I. Schiff, Am. J. Phys. **22**, 621, 1954, with permission.)

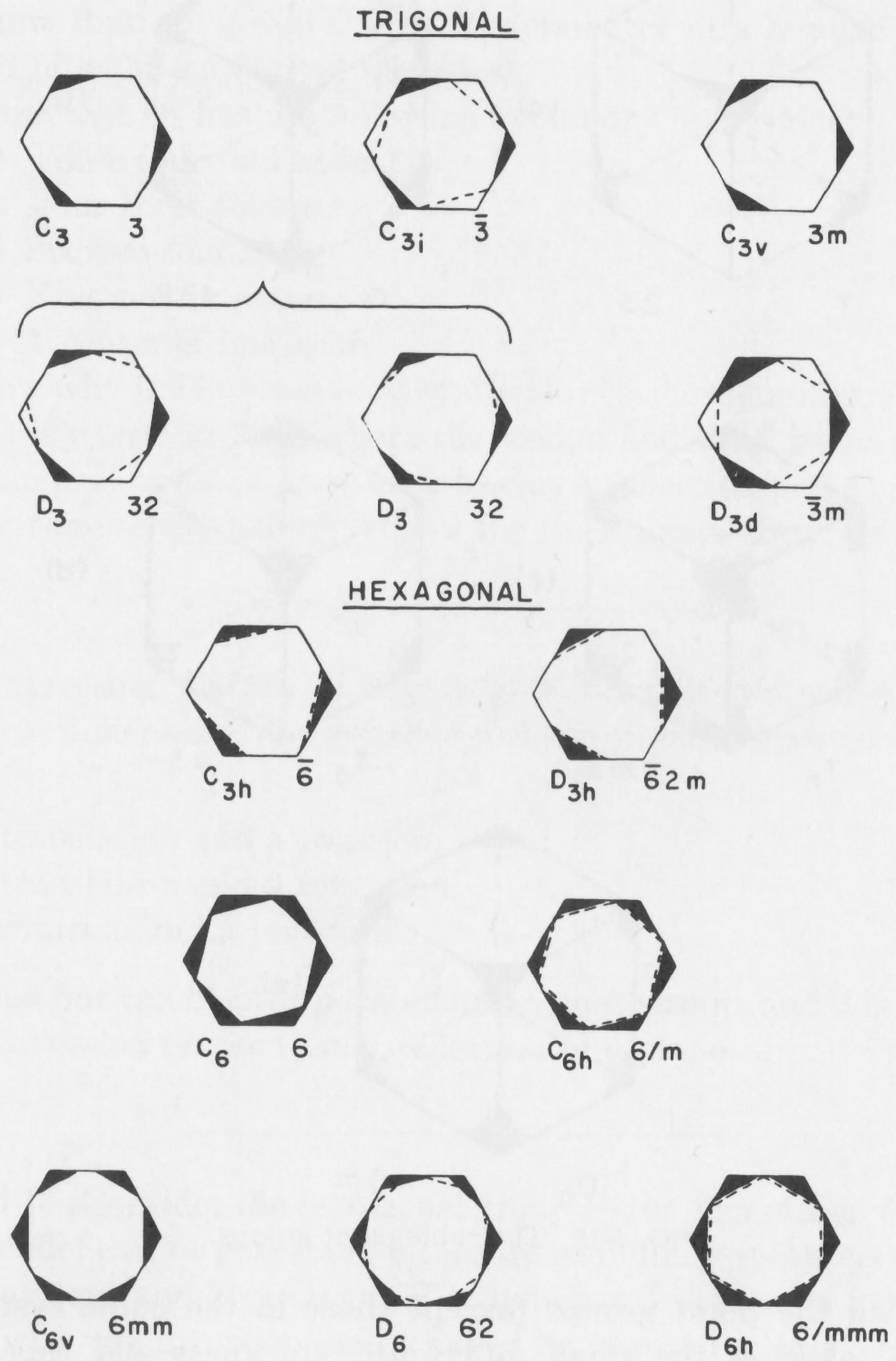

TRIGONAL
C_3 3
C_{3i} $\bar{3}$
C_{3v} 3m
D_3 32
D_3 32
D_{3d} $\bar{3}m$
HEXAGONAL
C_{3h} $\bar{6}$
D_{3h} $\bar{6}2m$
C_6 6
C_{6h} 6/m
C_{6v} 6mm
D_6 62
D_{6h} 6/mmm

FIG. 5-18(Continued).

The next group, D_2 or 222, is called a *dihedral* group; it denotes the existence of several two-fold axes perpendicular to a principal n-fold axis.

PROBLEM 5-9

(a) Explain why the notation is 222 rather than 22.

(b) For the trigonal dihedral group D_3 or 32, show that there are three two-fold axes perpendicular to the principal three-fold axis.

Rather than complete the enumeration in the above fashion, we shall use a clever method devised by Schiff (1954).[3] Figure 5-18 represents cardboard

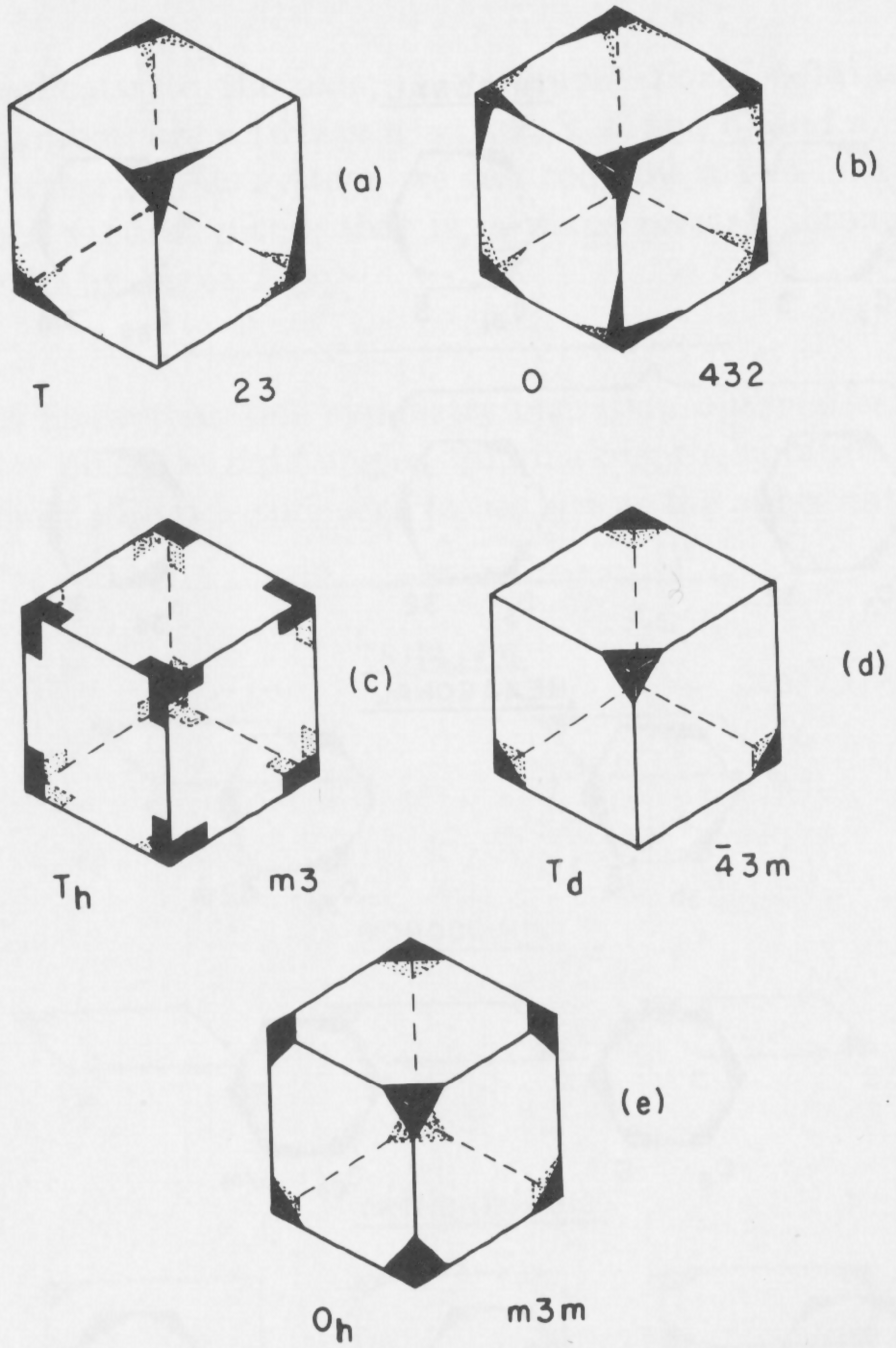


FIG. 5-19. The cubic point groups.

models of all the point groups (except those in the cubic system). Advantage is taken of the small, but finite, thickness and the symmetry properties are illustrated by the operations which can be performed on these models which leave the shaded triangles in positions similar to their original ones. (The dotted lines indicate triangles on the bottom surface.) Two representations of the class D_3 or 32 are given, for the two-fold axes can be perpendicular to the edges of the hexagon or can pass through the corners. For the cubic groups, it is necessary to use the drawings of Fig. 5-19 to show the symmetry properties.

PROBLEM 5-10

(a) Show that the point group T has the same symmetry as a regular tetrahedron (a figure composed of four equilateral triangles).

(b) Show that the group O has the symmetry of a regular octahedron (composed of eight equilateral triangles).

(c) Show that O_h has the following symmetry operations:

(1) Three four-fold axes.
(2) Four three-fold axes.
(3) Six two-fold axes.
(4) Nine mirror planes.
(5) A center of inversion.

(d) Show why the international symbol for the hexagonal Bravais lattice (Fig. 5-15) is $6/m\ 2/m\ 2/m$, where the second and third items denote two different kinds of two-fold axes, each having a reflection plane normal to it. Show why the corresponding symbol for the trigonal Bravais lattice is $\bar{3}\ 2/m$.

One interesting feature of Fig. 5-18 is brought out when we realize that in three dimensions it is possible to have three compound operations, as follows:

(1) A translation and a rotation.
(2) A translation and a reflection.
(3) A rotation and a reflection.

Only the last one can be incorporated into a point group, and it is customary to use an inversion rather than a reflection in a plane.

PROBLEM 5-11 Consider the tetragonal group D_{2d} or $\bar{4}2m$ of Fig. 5-18. Show that this model can be generated either by combining rotations and reflections or rotations and inversions. The notation $\bar{4}$ denotes a four-fold axis combined with the inversion and is called an *alternating* axis. Explain the significance of the remainder of the symbol $\bar{4}2m$.

The concept of the alternating axis enables us to introduce an idea which is often overlooked in many discussions of crystal structure: the relations among the a_i and the α_i given in Section 5-3 do *not* define the seven crystal systems, but only specify their geometrical properties. For example, a crystal with $\alpha_1 = \alpha_2 = \alpha_3$ and $\alpha_1 = \alpha_2 = \alpha_3 = 90°$ is not necessarily cubic; it would only belong to this system if the arrangement of atoms at each lattice point were consistent with cubic symmetry. The minimum number of symmetry elements which a lattice must possess to belong to a given system can be deduced from Fig. 5-18, and are as follows:

System	*Minimum symmetry*
Triclinic	None
Monoclinic	A two-fold axis
Orthorhombic	Two orthogonal two-fold axes or two orthogonal mirror planes
Tetragonal	A four-fold or a four-fold alternating axis
Cubic	Four three-fold axes arranged like the diagonals of a cube
Trigonal	A three-fold or a three-fold alternating axis
Hexagonal	A six-fold or a six-fold alternating axis

PROBLEM 5-12 Verify that the systems of Fig. 5-15 and the point groups of Figs. 5-18 and 5-19 satisfy the above conditions.

We have now surveyed the 32 crystal point groups, but we have not yet shown that no others are possible. Let us look into the question of how many symmetry axes can belong to any one point group. As an example of the limitations on the number and kinds of axes, consider two two-fold axes in a given group. Let these two axes be inclined at an angle of α to each other, as shown in Fig. 5-20. A lattice point P is transformed into P' by one of the two-fold rotations, and to P'' by the other, where OP, OP', and OP'' are all in the same plane. The total rotation γ of OP is then

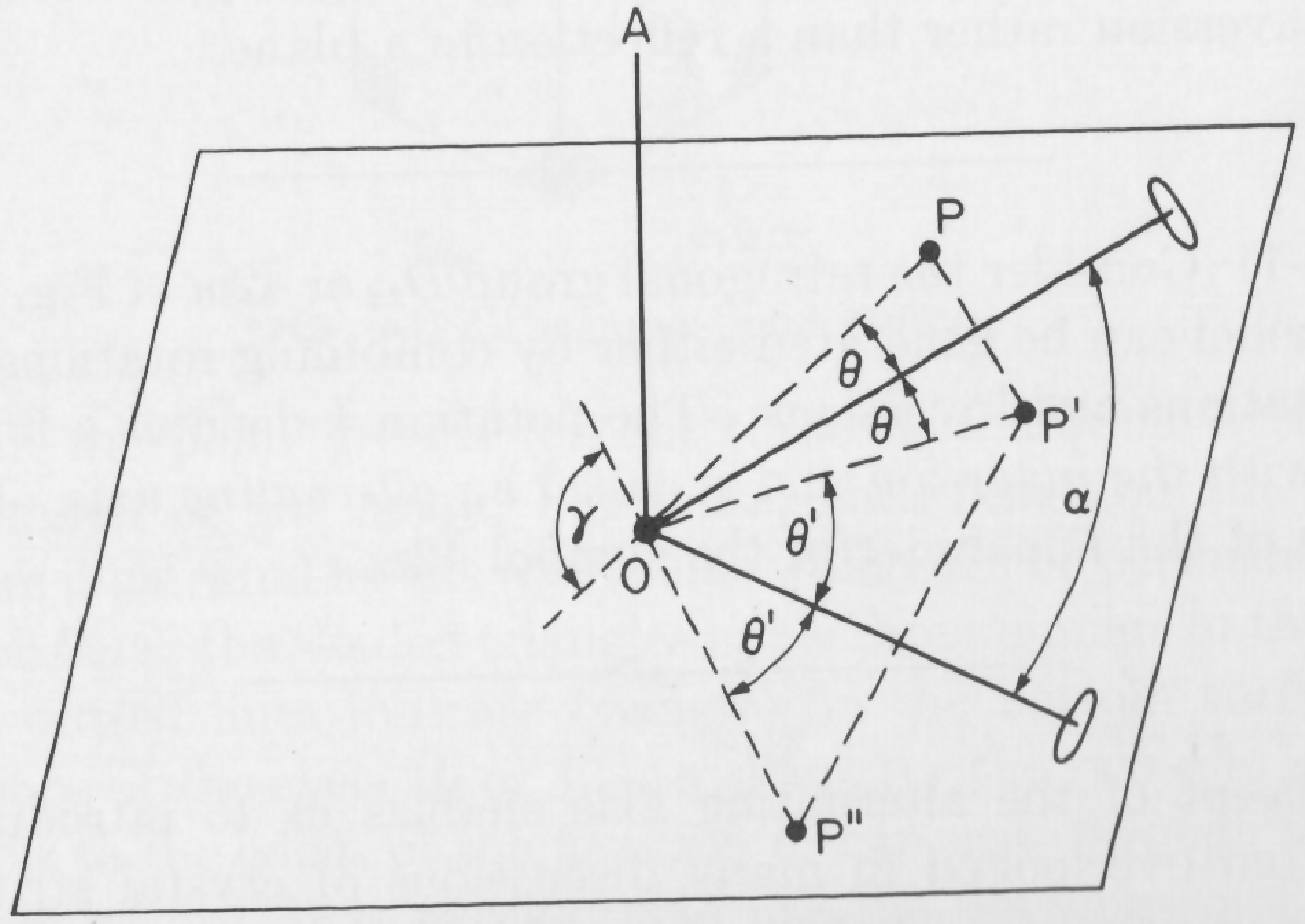


FIG. 5-20. A diagram used to show that the angle α between a pair of two-fold axes is restricted to certain values. (See text.)

$$\gamma = \theta + \theta + (\alpha - \theta) + (\alpha - \theta) = 2\alpha$$

Now, 2α must equal $2\pi/n$ where $n = 1, 2, 3, 4$ or 6. Hence,

$$\alpha = 180°, 90°, 60°, 45°, \text{ or } 30°$$

and no other inclinations of two two-fold axes are possible.

PROBLEM 5-13

(a) Show that two six-fold axes cannot exist in a point group.

(b) By considering the first five entries of Table 5-4, show that a *biclinic* lattice is inconsistent with any of the point groups.

5-5 The Space Groups

It will be recalled that the seventeen plane groups, examples of which are given in Fig. 5-13, were formed as follows: thirteen of them came by combining the ten plane point groups with five plane lattices, and the other four came when glide operations were introduced. The situation in three dimensions is analogous to this, but far more complicated. It is found that the thirty-two point groups, when combined in all the ways which are consistent with the fourteen Bravais lattices, give a total of 73 combinations, called *simple space groups*. We have already shown, in Fig. 5-17, one of the two triclinic space groups. Let us recall two of the compound symmetry operations which were introduced in Section 5-4, but were not discussed there because they did not belong to operations about a point.

The combination of a translation and a reflection in a plane is a *glide plane* and a translation combined with an n-fold rotation axis is a *screw axis*. The glide plane is analogous to the glide line of Fig. 5-11, and the screw axis is illustrated in Fig. 5-21, where an example of Phillips (1946) is used. Figure 5-21a shows a three-fold right-handed screw axis; a 120° turn to the right followed by a translation $a/3$ along the axis, where a is the lattice constant, brings the lattice back into coincidence with itself. In Fig. 5-21b, which represents a left-handed screw axis, a rotation in the opposite sense plus a translation $a/3$, or a right-handed rotation plus a translation of $2a/3$, gives the same result. Taking the right-handed operation as standard accounts for the international symbols 3_1 and 3_2.

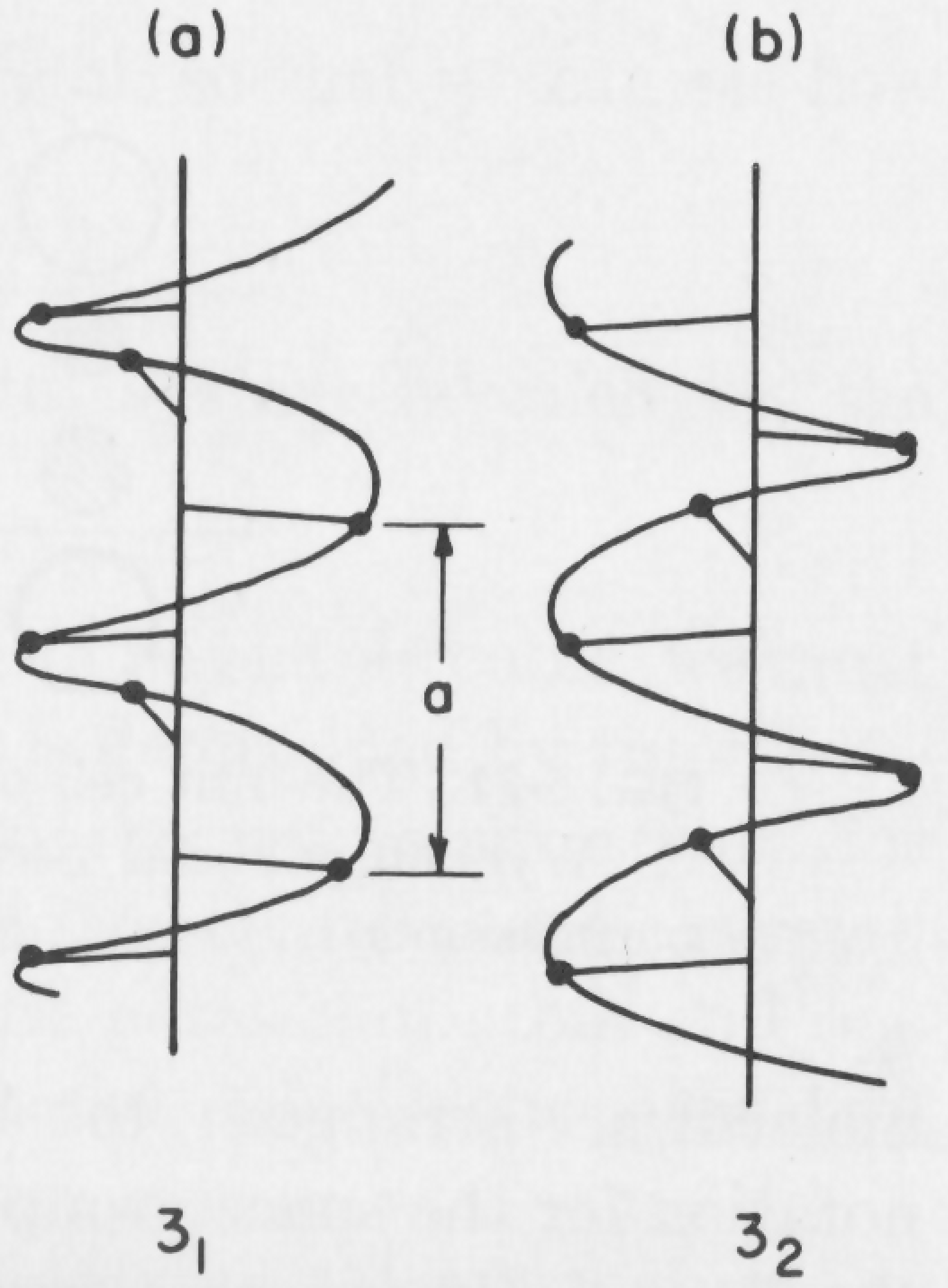


FIG. 5-21. Three-fold (a) right-handed and (b) left-handed screw axes. (From F. C. Phillips, *An Introduction to Crystallography*, Longmans, Green, London, 1946, with permission.)

Including the screw-axis and glide-plane as symmetry elements, it is possible to construct an additional 157 *compound space groups*, making a total of 230 space groups. These groups are enumerated and described in the article of Jagodzinski (1955)[4] and the introductory book of Phillips (1946).[5] It will suffice for our understanding of semiconductors to consider just a few examples here, rather than go into detail on all 230 groups. It should be borne in mind that a space group is not a crystal; a real crystal is obtained by placing an atom or a group of atoms at every point in the lattice. As a reasonably complex example, consider lithium sulfate monohydrate, $Li_2SO_4 \cdot H_2O$. According to Wyckoff (1960),[6] this compound crystallizes in the monoclinic form, as shown in Fig. 5-22. The unit cell has two $Li_2SO_4 \cdot H_2O$ molecules, arranged to have two-fold symmetry. The Schoenflies notation for the space group is C_2^1, where the superscript means that this is the first of fifteen space groups to have this symmetry; the abbreviated international symbol is $P2$. Note that the detailed structure of the water molecule is not shown; the determination of crystal structure is done by the use of x-ray methods, and it would take extremely refined methods to determine the exact orientation of the atoms in the H_2O molecule.

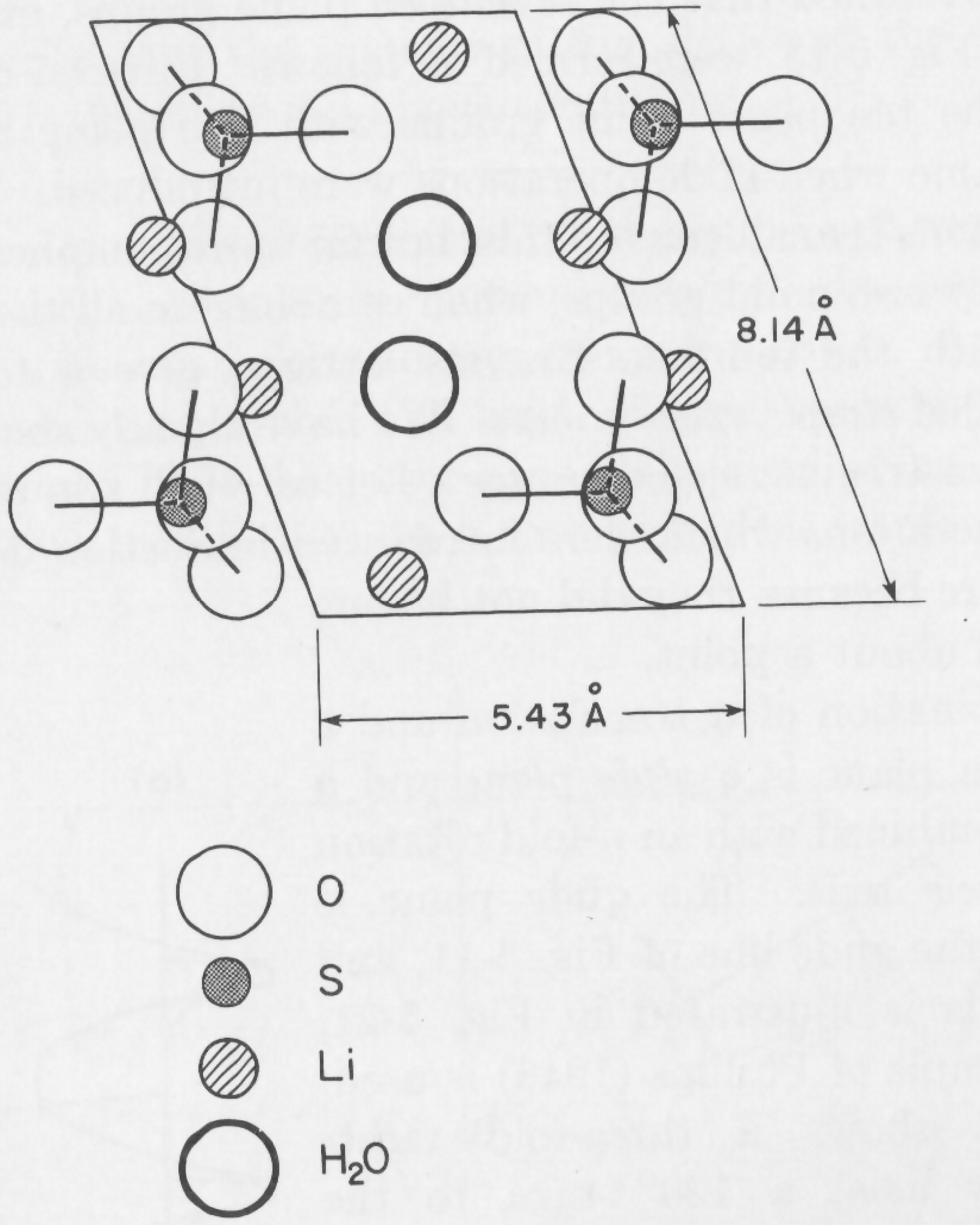


FIG. 5-22. The unit cell of lithium sulfate monohydrate. (After R. W. G. Wyckoff, *Crystal Structures*, Interscience, New York, 1960, with permission.)

Let us now turn to the cubic lattices, where the most useful semiconductors are found. As a very simple example, consider *sylvine*, potassium chloride, whose lattice is shown in Fig. 5-23. The dotted lines indicate the shape of an occasional perfect natural crystal, so that we can see how this shape originates.

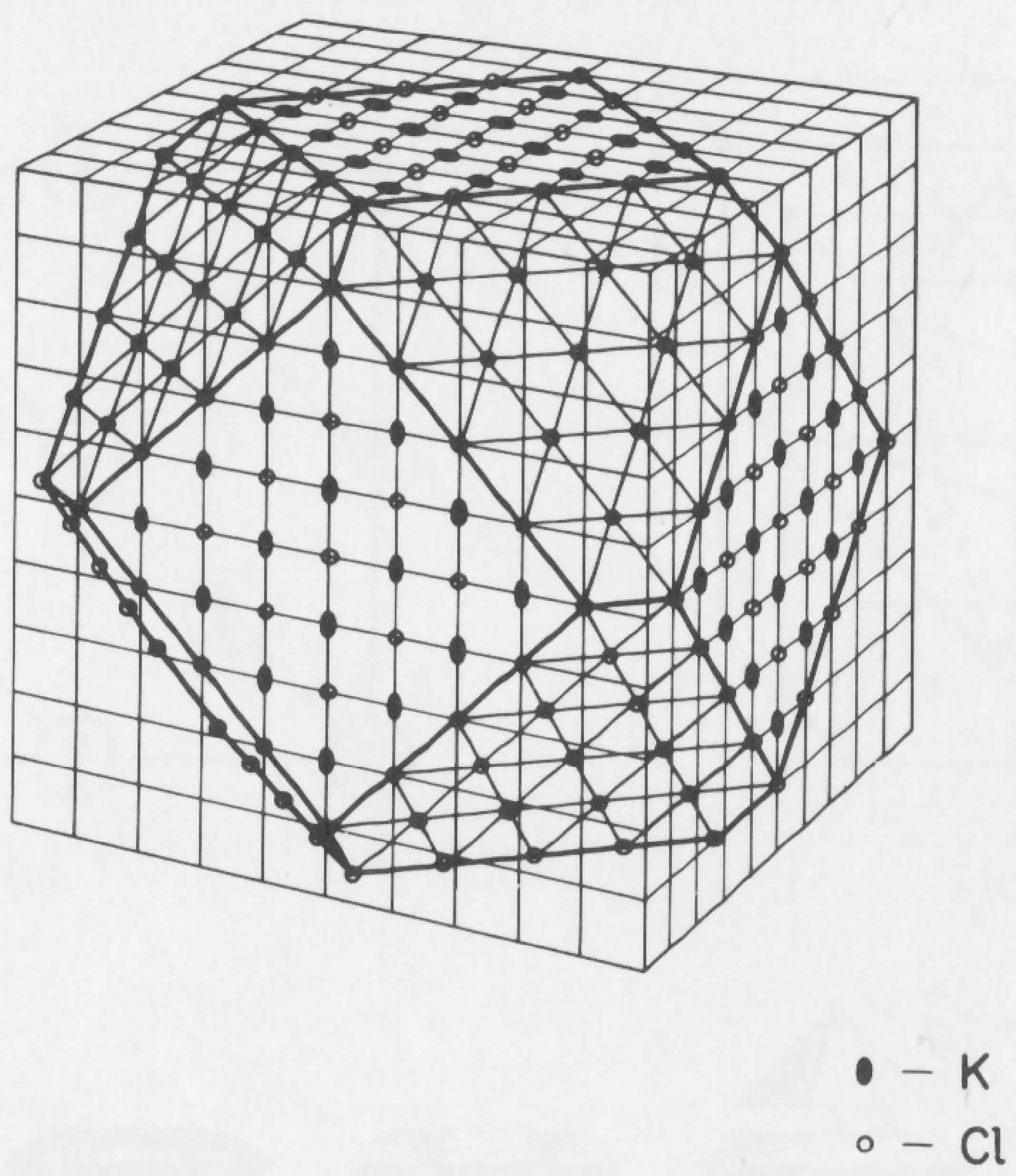
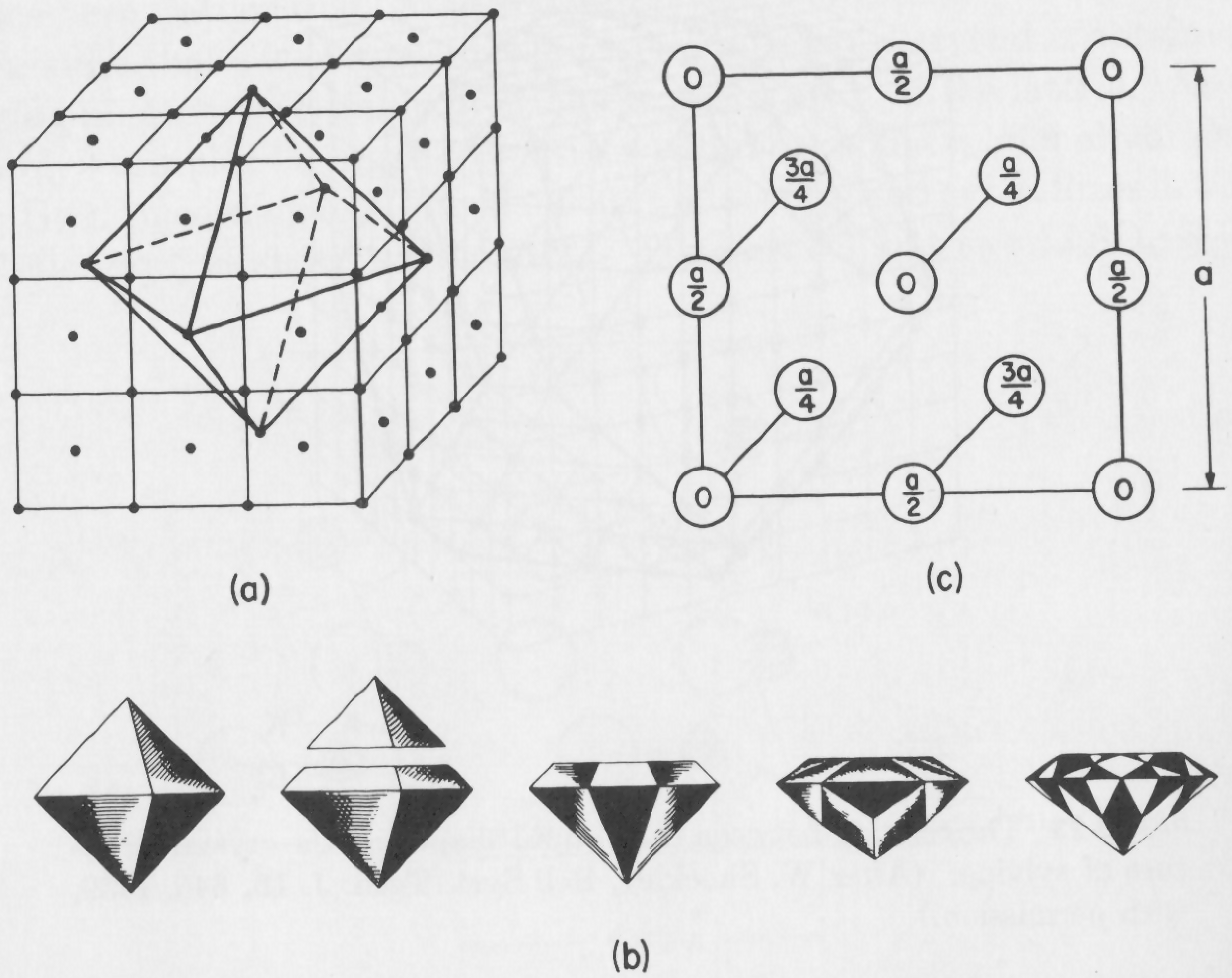


FIG. 5-23. The relation between the natural shape and the crystal structure of sylvine. (After W. Shockley, Bell Syst. Tech. J. **18,** 645, 1939, with permission.)

PROBLEM 5-14 Is the symmetry of the crystal system revealed by the natural crystal?

Although the lattice appears at first glance to be simple cubic, we must bear in mind that the translational symmetry is determined by the repetition of a given atom, and if we consider just the potassium atoms, for example, then we see that Fig. 5-23 illustrates a face-centered cubic crystal. The space group is constructed by placing one potassium atom and one chlorine atom at each point on the face-centered cubic lattice, the two atoms being separated by a distance $a/2$, where $a = a_1 = a_2 = a_3$.

The most important examples of the face-centered cubic structure are the set which we have mentioned several times before: diamond, silicon, germanium, and gray tin. The lattice is illustrated in Fig. 3-3, and Fig. 5-24a shows how the octahedral shape of uncut diamonds is determined by the atomic arrangement (Fig. 5-24b shows one of the methods of cutting and polishing diamonds to get the common 58-facet type). The diamond lattice is constructed by placing two atoms at each point, one of them being displaced by a distance $a/4$ along each of the three axes. This structure can be visualized by referring to Fig. 5-24c, in which the atoms of the original face-centered lattice are shown joined by lines to the corresponding atoms of the displaced lattice, and the heights of the atoms above the XOY plane are denoted by the quantities inside the circles.



(a) (c)

(b)

FIG. 5-24. The diamond lattice, showing (a) the origin of the octahedral shape, (b) a method of cutting the gems (reproduced with permission of N. W. Ayer & Son, Philadelphia), and (c) the unit cell. (a is the lattice constant and the fractions in the circle denote the height above the base plane.)

PROBLEM 5-15 Show that the space group of Fig. 5-24c has all the symmetry elements of the point group O_h listed in Problem 5-10(c).

The notation for the space group of the diamond lattice is O_h^7 (the superscript denoting that this is the seventh of ten lattices with O_h symmetry) or $F\ 4_1/d\ \bar{3}\ 2/m$, abbreviated $Fd3m$. The F is for face-centered; the 4_1 denotes a four-fold screw axis, which is indicated by the numbers 1 through 5 of Fig. 3-3; and the d denotes a *diagonal glide*.

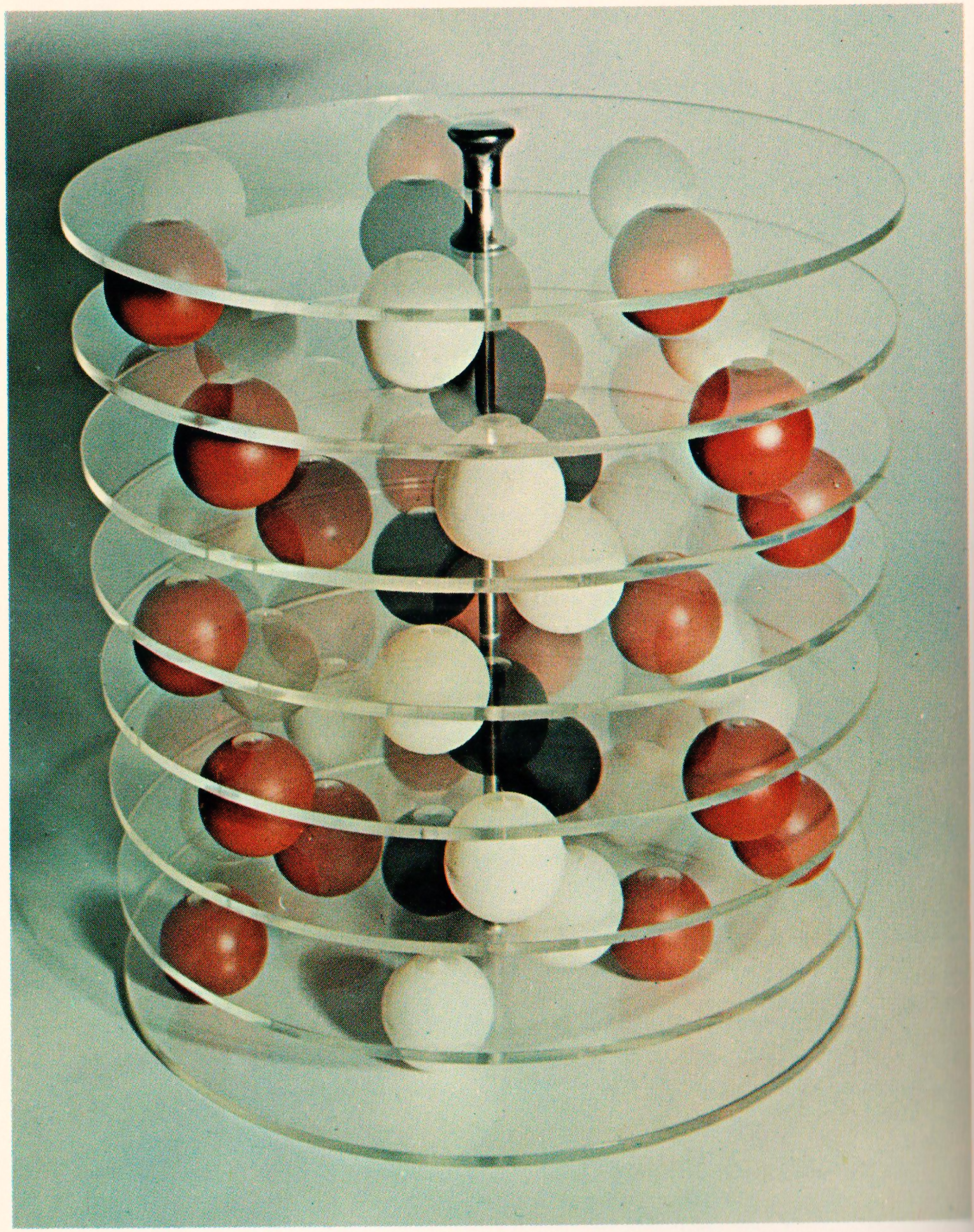


FIG. 5-25 (b). A scale model of the tellurium lattice. Seven equivalent spiral chains are shown in contrasting colors as an aid to visualizing the structure.

PROBLEM 5-16

(a) In Fig. 5-24c, locate one of the screw-axes, and show that every atom in the figure will move into coincidence with another atom under the screw operation.

(b) Let the plane of the paper be the XOY plane with the origin in the lower left-hand corner, and let m_x denote a reflection in the plane $x = \frac{1}{4}$. The diagonal glide is then defined as the operation which consists of the displacement $a/4$ in the OY and OZ directions followed by the reflection m_x. Show that this operation brings the lattice into coincidence with itself.

The cubic groups mentioned in Problem 5-10 can be used to illustrate an idea which has not as yet been explicitly covered. Of the five groups T, T_h, T_d, O, and O_h in Table 5-1, the most symmetrical one is O_h, which has a total of 48 symmetry operations. In working out Problem 5-10c it will be seen that these operations can be enumerated in the following way:

1. The simplest operation which leaves a cube coincident with itself is not to move it at all, and this is called the *identity*.
2. Considering just one of the four-fold axes, the three distinct symmetry operations are a 90° clockwise rotation, a 90° counter-clockwise rotation, and a 180° rotation (in either direction). Hence, there are nine possible operations for the three axes.
3. Similarly, there are eight operations associated with the three-fold axes and six with the two-fold axes.
4. The inversion by itself produces the twenty-fifth operation, and when combined with (2) and (3) above, the final total is forty-eight.

This point group O_h of 48 operations is based on the properties of the cubic translational lattice and is called a *holosymmetric* group. If we place a single atom at every point in the lattice, the resulting crystal will have the same high symmetry, but we already know that clusters of atoms of lower symmetry can be placed at each point. Hence it is possible to have cubic crystals with only twenty-four associated symmetry operations.

As our final example of space groups, we shall consider the elements selenium, tellurium, and polonium occurring in the lower part of column VIA. Figure 5-25a shows a sketch of a tellurium crystal grown by T. J. Davies (1957)[7]; Fig. 5-26b shows a scale model of this crystal; Fig. 5-26c is an end-on view, showing the three-fold screw axes. As we have already mentioned, the screw axes can be right-handed or left-handed, and the symbols for the corresponding space groups are D_3^4 or $P3_121$ and D_3^6 or $P3_221$.

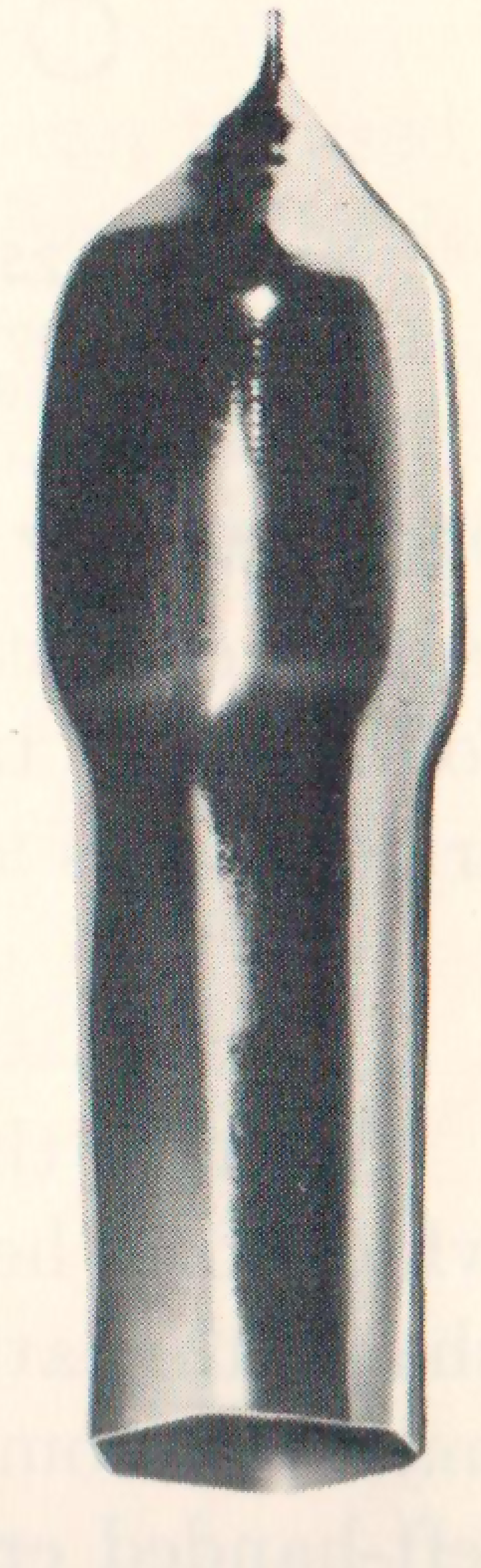
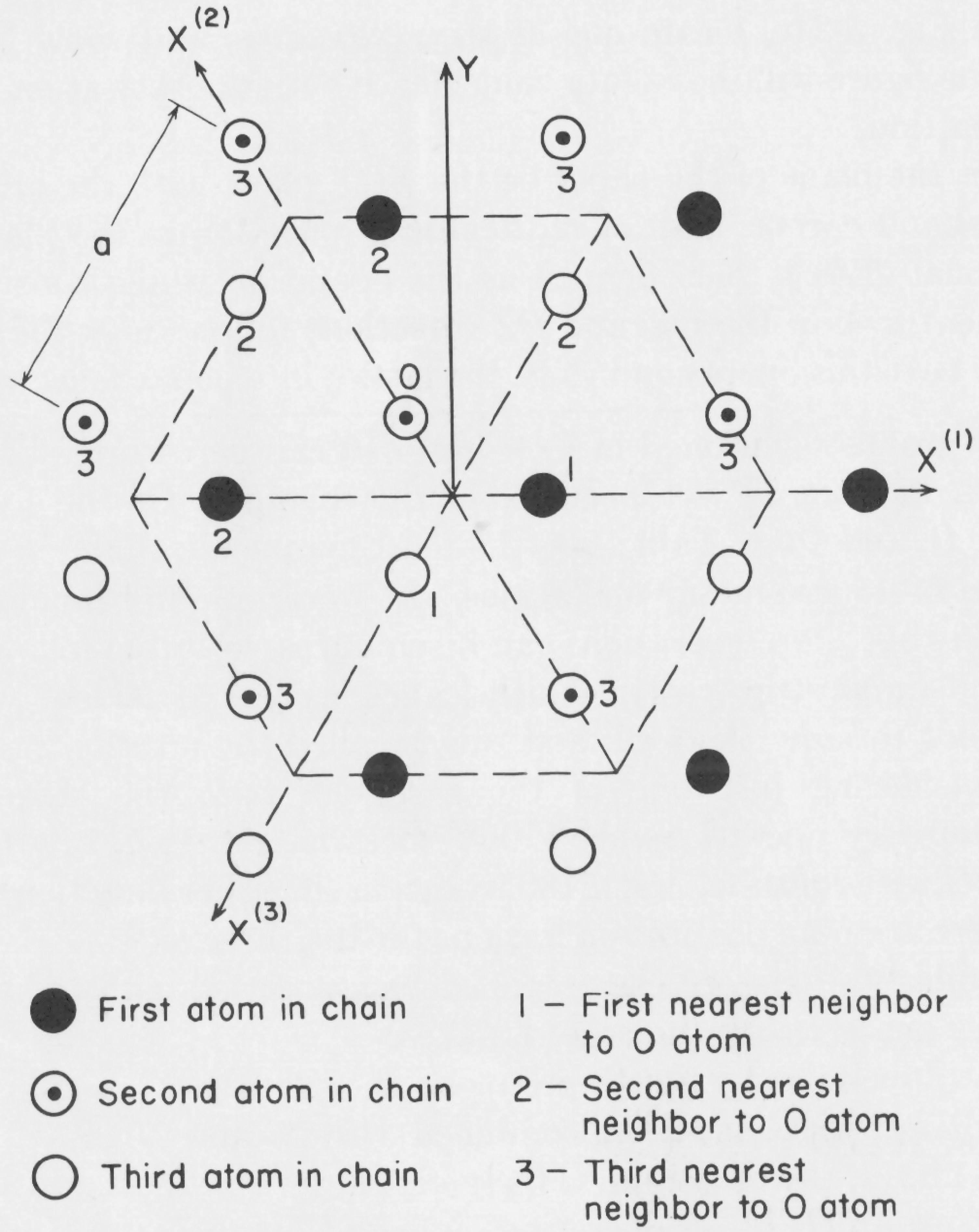


FIG. 5-25 (a). A tellurium crystal (grown by T. J. Davies.)



X(2)
Y
X(1)
X(3)
a
0
1
2
3
First atom in chain
Second atom in chain
Third atom in chain
1 – First nearest neighbor to 0 atom
2 – Second nearest neighbor to 0 atom
3 – Third nearest neighbor to 0 atom

FIG. 5-25 (c). The tellurium lattice. (From R. N. Asendorf, J. Chem. Phys. *27*, 11, 1957, with permission.)

PROBLEM 5-17 Show that three of the symmetry operations of the tellurium lattice consist of 180-deg rotations about the axes $X^{(1)}$, $X^{(2)}$, or $X^{(3)}$ followed by translations $nc/3$, where c is the lattice constant along the trigonal axis and $n = 0$, 1, or 2.

We note that, while the atoms in tellurium are arranged in spiral chains which form hexagonal bundles, and the crystal itself shows the hexagonal shape, the lattice symmetry is trigonal, due to the fact that three atoms make one complete cycle in the spirals. The existence of right-handed and left-handed crystals was discovered by Nomura (1960).[8]

Polonium, which is a radioactive metal, has been reported by Beamer and Maxwell (1946)[9] to be simple cubic at room temperature, making it the

only such crystal known. The relation between the simple cubic structure and the trigonal chains of tellurium and selenium was first noticed by Bradley (1924).[10] He showed, as we see in Fig. 5-26, that the displacement

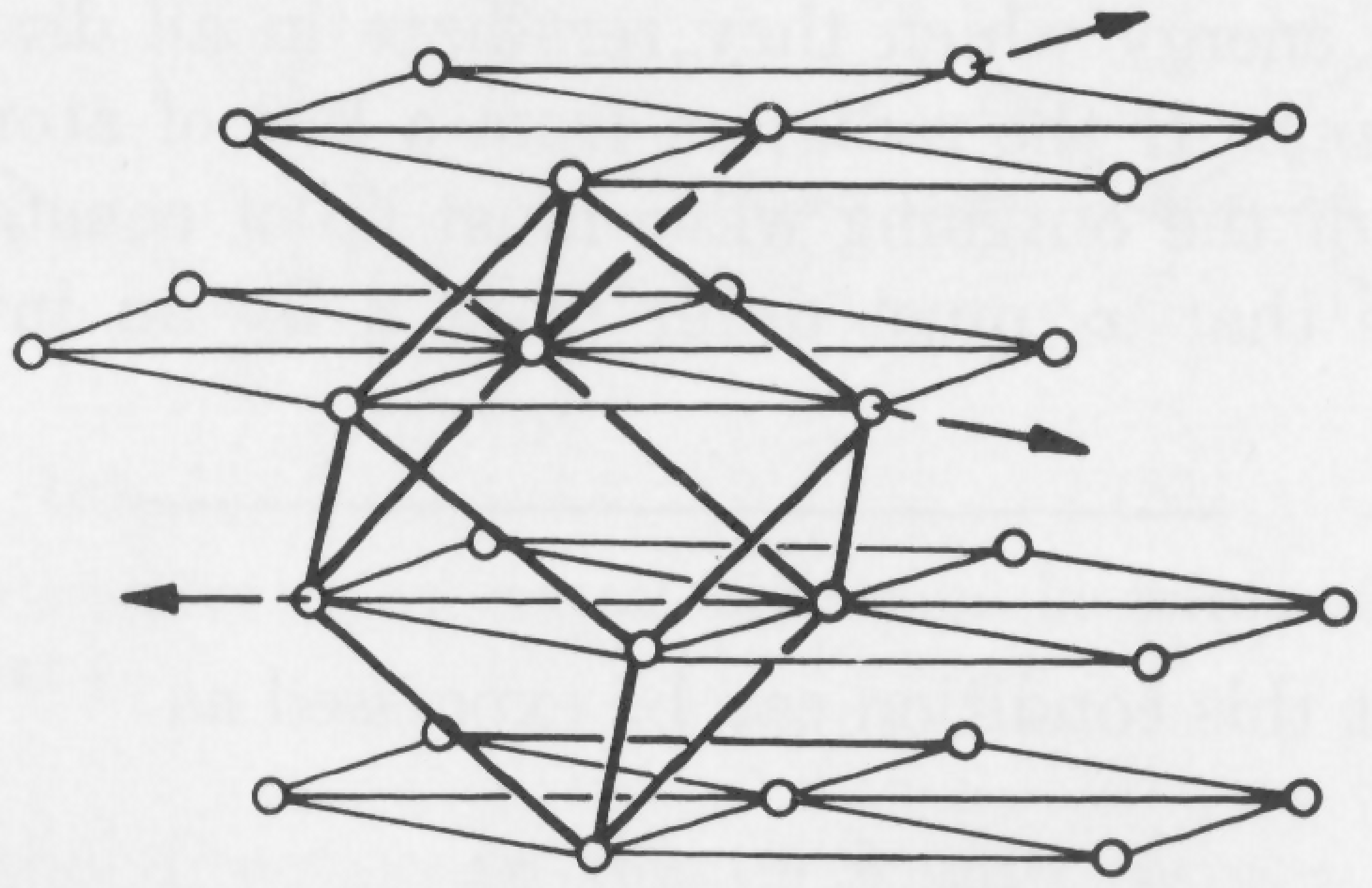


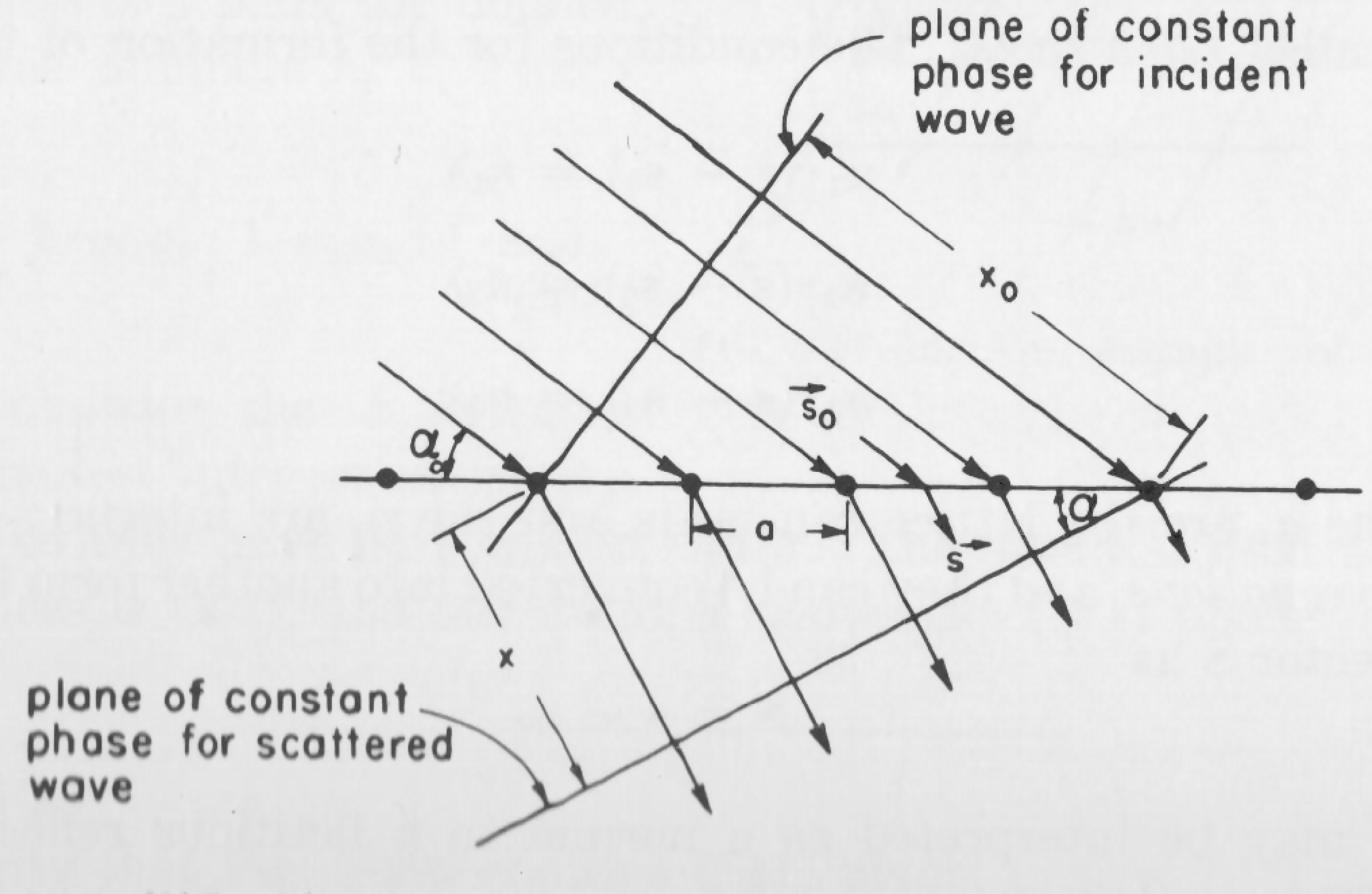
FIG. 5-26. The relation of the simple cubic structure of polonium to that of tellurium. (From A. R. von Hippel, J. Chem. Phys. *16*, 372, 1948, with permission.)

of three adjacent planes of polonium along directions which are 120° to each other produces the screw-axis structure of tellurium or selenium.

PROBLEM 5-18 By sketching a top view of the lattice in Fig. 5-26, show that it can be converted into Fig. 5-25c.

5-6 Crystal Structure and Brillouin Zones

It was mentioned in the previous section that x-rays can be used to determine the nature of crystal lattices. We shall now discuss the laws



plane of constant phase for incident wave
x₀
s₀
α₀
a
α
s
x
plane of constant phase for scattered wave

FIG. 5-27. Diffraction by a crystal. (From A. J. Dekker, *Solid State Physics*, Prentice-Hall, Englewood Cliffs, N. J., 1957, with permission.)

governing the diffraction of x-rays by crystals, since this provides an opportunity of introducing some ideas which are basic to the study of energy band theory. Consider a plane-wave x-ray striking a lattice at an inclination α_0, as shown in Fig. 5-27. When x-rays strike an atom, the electrons receive energy which they reradiate in all directions, a process known as *scattering.* If the radiation from a line of atoms is to reinforce itself, the plane of the outgoing wave must be of constant phase. Thus, the figure shows that x_0 must differ from x by an integral number of wave lengths.

PROBLEM 5-19

(a) Show that this condition can be expressed as

$$x - x_0 = n\lambda$$

where $n = 0, 1, 2, \ldots$, using the interpretation of phase developed in Problem 1-3.

(b) Show that in vector form, this equation becomes

$$\mathbf{a}\cdot(\mathbf{s} - \mathbf{s}_0) = n\lambda$$

where $\mathbf{s}_0$ and $\mathbf{s}$ are unit vectors along the directions of the incident and scattered waves, respectively.

It should be realized that the scattered waves will reinforce over a complete cone of apex half-angle α around the line of atoms. Hence, when we consider a two-dimensional lattice, we get reinforcement only on the lines where the cones meet; in three dimensions, the intersections are at points, called *Laue spots.* The conditions for the formation of these spots are

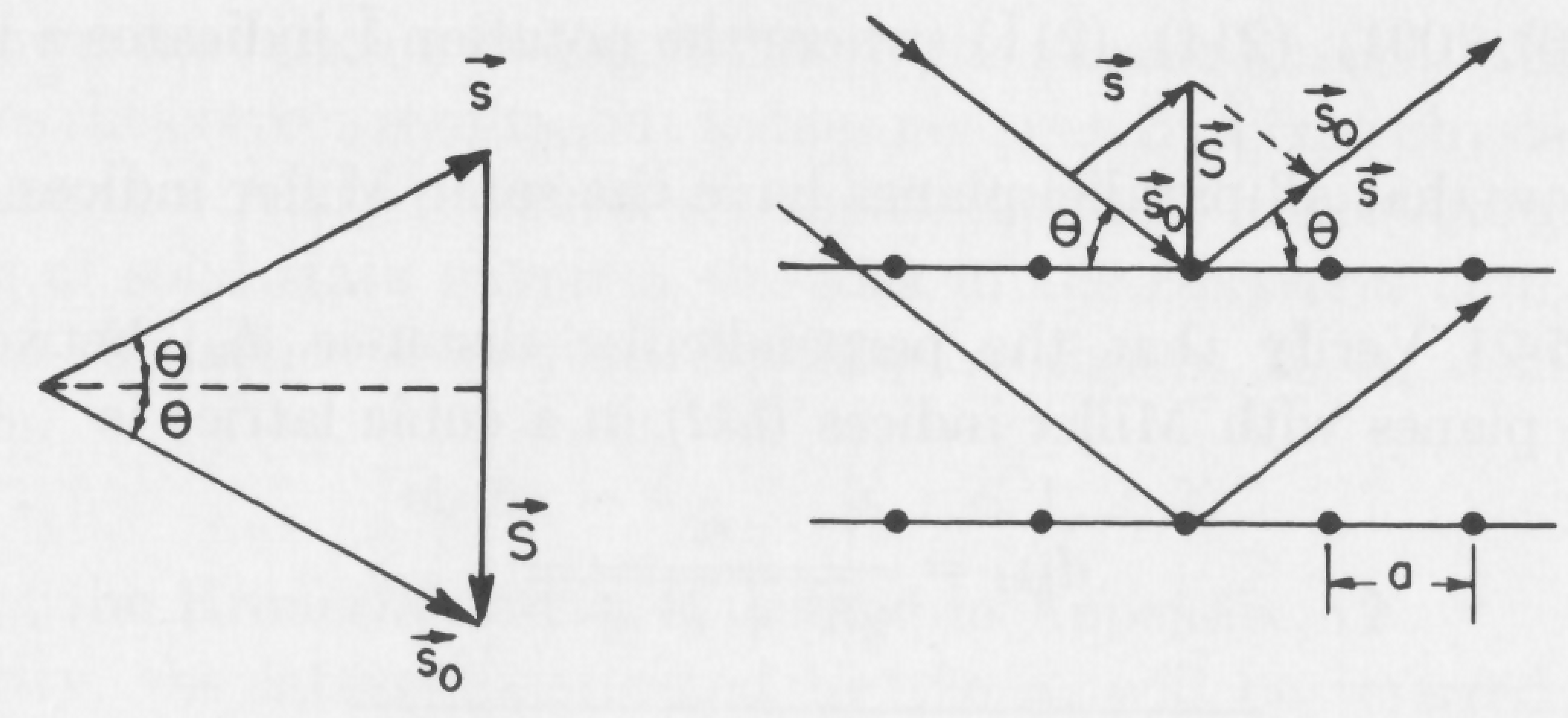
$$\begin{aligned} \mathbf{a}_1\cdot(\mathbf{s} - \mathbf{s}_0) &= n_1\lambda \\ \mathbf{a}_2\cdot(\mathbf{s} - \mathbf{s}_0) &= n_2\lambda \qquad (5\text{-}2) \\ \mathbf{a}_3\cdot(\mathbf{s} - \mathbf{s}_0) &= n_3\lambda \end{aligned}$$

where the a_i are the lattice constants and the n_i are integers. These are the *Laue equations,* and they can be converted into another form by defining a new vector $\mathbf{S}$ as

$$\mathbf{S} = \mathbf{s} - \mathbf{s}_0$$

Then $\mathbf{S}$ may be interpreted as a normal to a fictitious reflecting plane (Fig. 5-28a), so that

$$S = 2s \sin\theta = 2s_0 \sin\theta \qquad (5\text{-}3)$$



(a) (b)

FIG. 5-28. Interpretation of crystal diffraction in terms of the Bragg law. (See text.)

If α_1, α_2, α_3 are the direction-cosines of $\mathbf{S}$ with respect to $\mathbf{a}_1$, $\mathbf{a}_2$, $\mathbf{a}_3$, respectively, then Eqs. (5-2) and (5-3) combine to give

$$\mathbf{a}_i\cdot\mathbf{S} = 2a_i\alpha_i \sin\theta = n_i\lambda \qquad (i = 1, 2, 3) \qquad (5\text{-}4)$$

To make the algebra easier, let us restrict the rest of this development to a simple cubic lattice and consider an arbitrary plane passing through a large number of atoms and striking the coordinate axes at points m_1a, m_2a, and m_3a, where the m_i have to be integers (Fig. 5-29a). In the example shown, the three intercepts are all equal to $3a$. The *Miller indices* of a plane are defined as the three numbers h, k, l such that

$$h : k : l = 1/m_1a_1 : 1/m_2a_2 : 1/m_3a_3 \qquad (5\text{-}5)$$

with the condition that h, k, and l are the smallest integers satisfying Eq. (5-5), so they have no common factor. The notation used to denote Miller indices is (hkl), and our example shows the (111) plane.

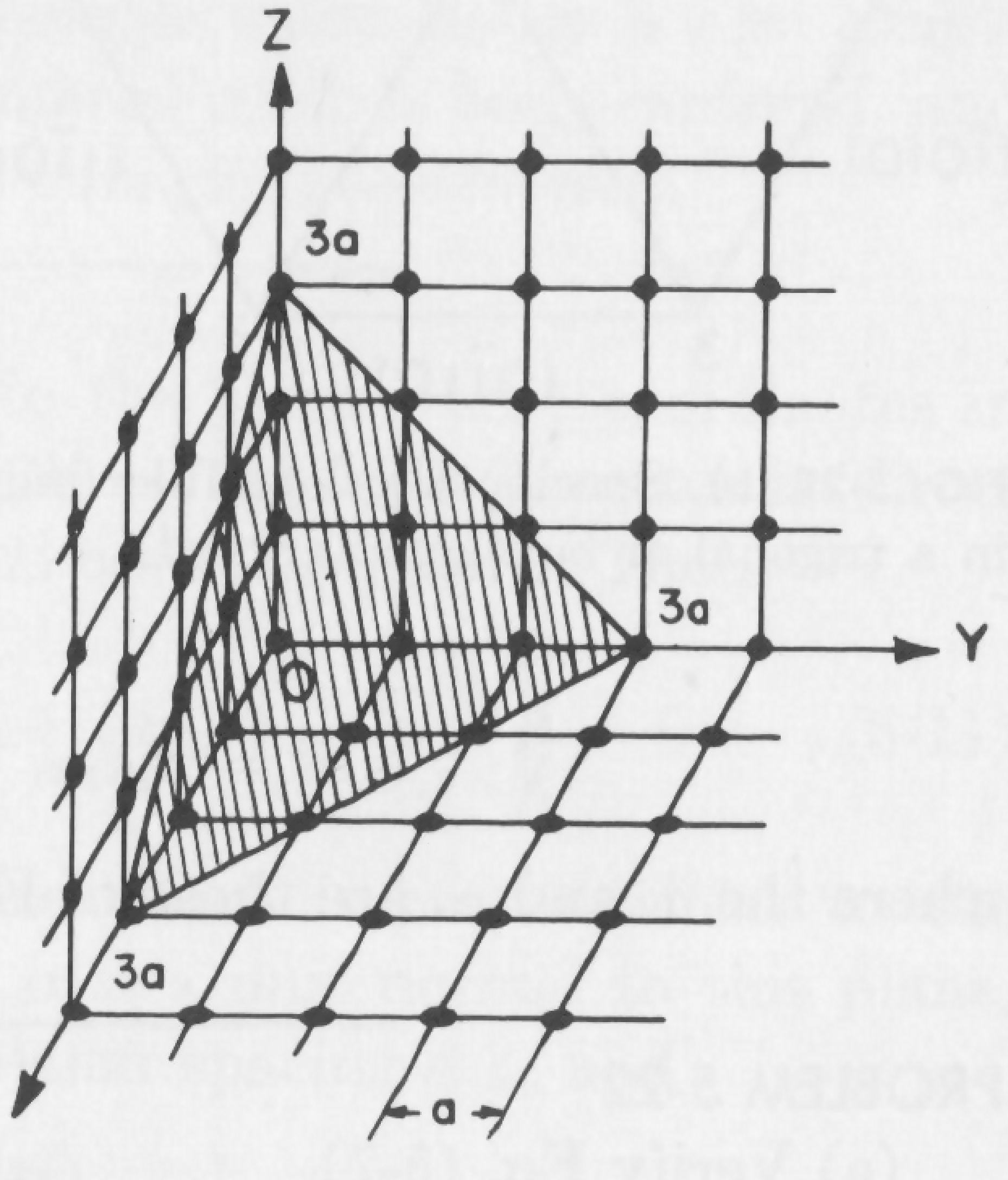


FIG. 5-29 (a). An example of a (111) plane.

PROBLEM 5-20

(a) Verify that Fig. 5-29a shows a (111) plane.

(b) Draw a similar diagram showing the following planes:

(100), (010), (001), (211), $(21\bar{1})$ (where the notation $\bar{1}$ indicates a negative intercept).

(c) Show that all parallel planes have the same Miller indices.

PROBLEM 5-21 Verify that the perpendicular distance d_{hkl} between two successive planes with Miller indices (hkl) in a cubic lattice is

$$d_{hkl} = \frac{a}{\sqrt{h^2 + k^2 + l^2}} \tag{5-6}$$

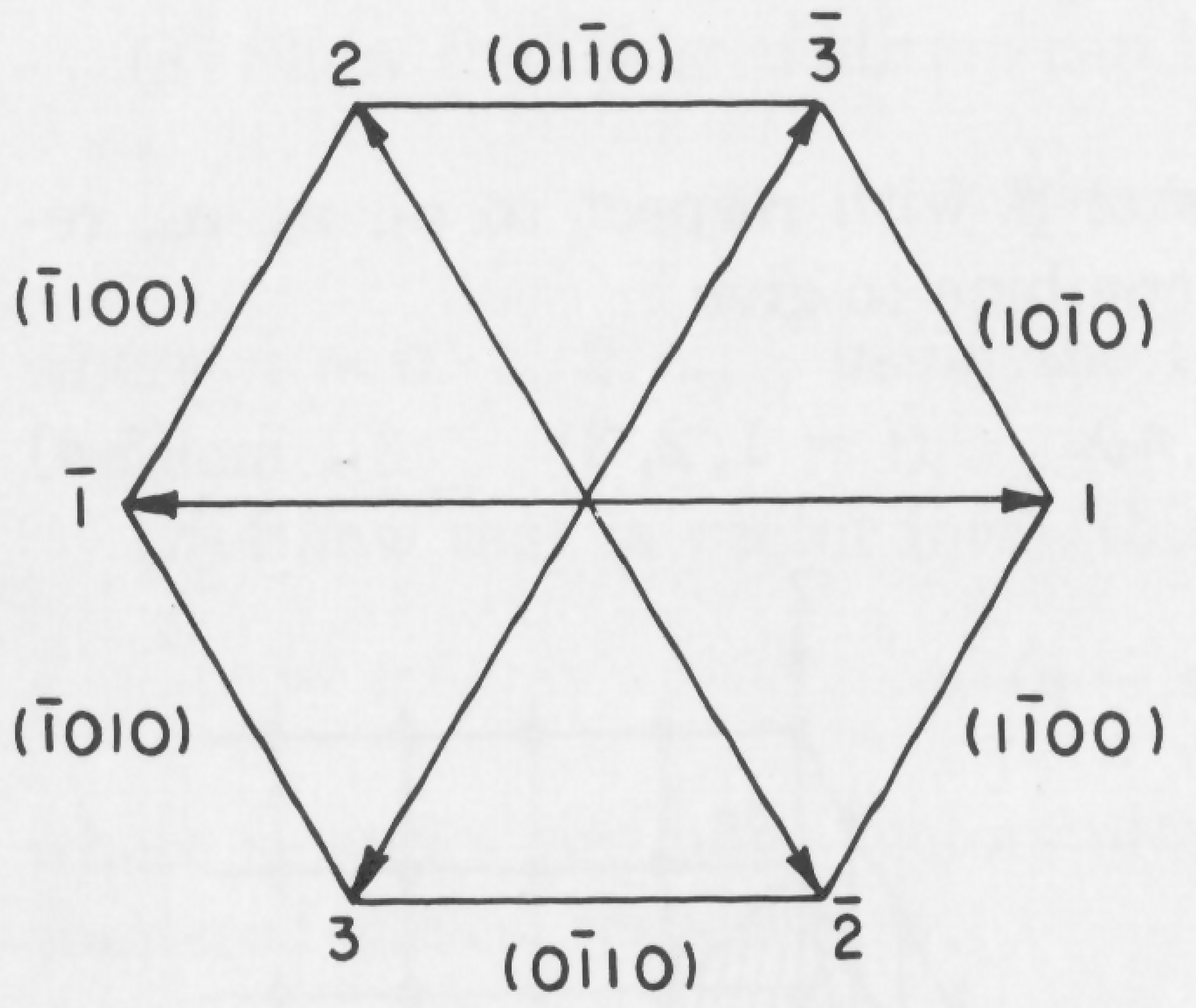


FIG. 5-29 (b). Specification of Miller indices in a trigonal or hexagonal crystal.

For the case of a trigonal or hexagonal crystal, such as tellurium, the specification of the Miller indices is more complicated. Figure 5-29b shows how the six faces of a hexagonal prism are indicated by the intercepts at each end of a face. It is seen that four integers are used; three for the XOY plane and the fourth to denote that all six faces are parallel to the z-axis.

We see from the reasoning used in Problem 5-21 that the spacing d between two adjacent parallel planes, one of which has intercepts n_1a_1, n_2a_2, n_3a_3, is

$$d = n_1a_1\alpha_1 = n_2a_2\alpha_2 = n_3a_3\alpha_3 \tag{5-7}$$

where the a_i and α_i are those of Eq. (5-4).

PROBLEM 5-22

(a) Verify Eq. (5-7).

(b) Let $k_i = 1/n_i$, and let the k_i have a common factor n. Show that Eqs. (5-4) and (5-7) combine to give

$$2d_{hkl} \sin \theta = n\lambda \tag{5-8}$$

which is known as the *Bragg law.*

Dekker (1957)[11] has pointed out that the two relations, Eqs. (5-2) and (5-8), are equivalent; however, the Bragg law was derived on the assumption that the atomic planes reflect like a mirror, as shown in Fig. 5-28b, where we have shown the relation of Fig. 5-28a to the atomic planes. This assumption gives the correct results, but it does not present a true physical picture.

We are now ready to introduce one of the most important geometrical concepts of solid state physics: the idea of the *reciprocal lattice.* This is defined as the lattice determined by three vectors $\mathbf{b}_1$, $\mathbf{b}_2$, $\mathbf{b}_3$, defined by the equations

$$\mathbf{a}_i \cdot \mathbf{b}_j = \delta_{ij} \qquad (i, j = 1, 2, 3) \tag{5-9}$$

where δ_{ij}, the Kronecker delta, is defined in Appendix A2

For clarity, the lattice determined by the $\mathbf{a}_i$ will be referred to as the *direct lattice.*

PROBLEM 5-23

(a) From Eq. (5-9), show that $\mathbf{a}_1$ is perpendicular to the plane of $\mathbf{b}_2$ and $\mathbf{b}_3$, and find the corresponding relation for the other five vectors.

(b) Prove that

$$\mathbf{b}_1 = \mathbf{a}_2 \times \mathbf{a}_3/[\mathbf{a}_1\ \mathbf{a}_2\ \mathbf{a}_3] \tag{5-10}$$

where the denominator is the box product as defined in Appendix A-2.

(c) Show that the reciprocal of the simple cubic lattice is also simple cubic, that the reciprocal of the face-centered cubic is body-centered, and that the reciprocal of the body-centered cubic is face-centered.

Let $\mathbf{r}_b$ be a vector from the origin to the point whose coordinates in reciprocal space are (h, k, l). Also, $[(\mathbf{a}_1/h) - (\mathbf{a}_2/k)]$ is a vector in direct space lying in the plane whose Miller indices are (h, k, l). Then

$$\mathbf{r}_b \cdot \left(\frac{\mathbf{a}_1}{h} - \frac{\mathbf{a}_2}{k}\right) = (h\,\mathbf{b}_1 + k\,\mathbf{b}_2 + l\,\mathbf{b}_3) \cdot \left(\frac{\mathbf{a}_1}{h} - \frac{\mathbf{a}_2}{k}\right) = 0 \tag{5-11}$$

Hence, $\mathbf{r}_b$ is a vector in reciprocal space which is perpendicular to the (hkl)-plane in direct space. Further, if $\mathbf{n}$ is a unit normal to this plane, then $\mathbf{a}_1 \cdot \mathbf{n}/h$, by Eq. (5-7), is the inter-planar spacing d_{hkl}, and

$$d_{hkl} = \left(\frac{\mathbf{a}_1}{h}\right) \cdot \left(\frac{\mathbf{r}_b}{r_b}\right) = \frac{1}{r_b} \tag{5-12}$$

Wilson (1953)[12] has shown that the three Laue equations [Eq. (5-2)] can be combined into a single equation in reciprocal space.

PROBLEM 5-24 Show that the equations in Eq. (5-2) are equivalent to

$$(\mathbf{s} - \mathbf{s}_0) = \lambda(n_1\mathbf{b}_1 + n_2\mathbf{b}_2 + n_3\mathbf{b}_3) \tag{5-13}$$

The vector on the right-hand side of Eq. (5-13) is a reciprocal-lattice vector — that is, it connects the origin and some lattice point — and is designated by OS in Fig. 5-30. The figure shows the geometrical interpretation of Eq. (5-13) in reciprocal space: draw a vector $-\mathbf{s}_0/\lambda$ from O to the point R (which need not be a reciprocal lattice point). From R, draw a sphere of radius $1/\lambda$. If this sphere passes through S, the Bragg-Laue law is satisfied. Further, if we let

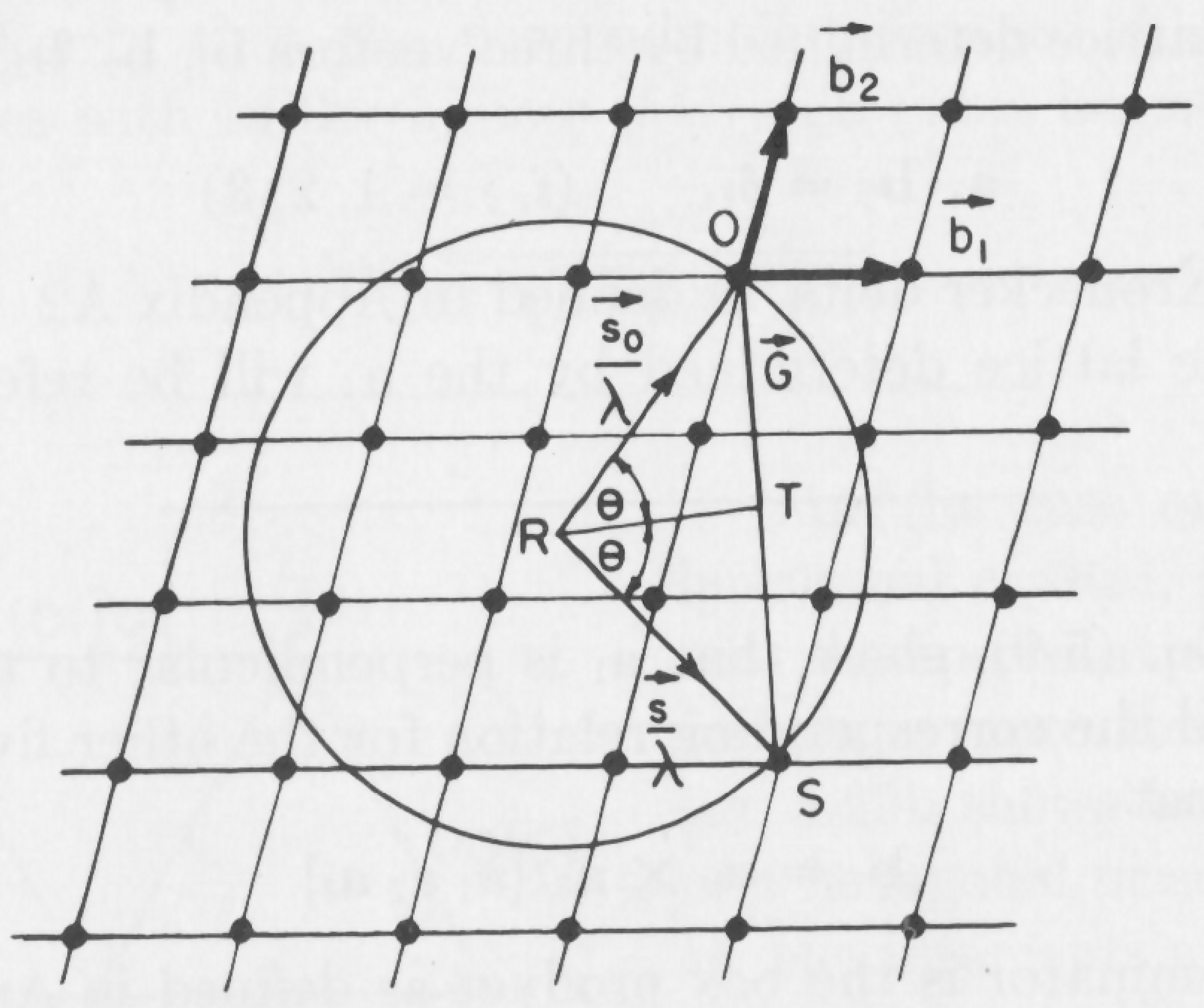


FIG. 5-30. Geometrical interpretation of the Bragg-Laue law in the reciprocal lattice. (From A. H. Wilson, *The Theory of Metals*, Cambridge Univ. Press, London, 1953, with permission.)

$$\mathbf{G} = 2\pi\, OS\, (\mathbf{s} - \mathbf{s}_0)$$

$$\mathbf{k} = 2\pi\, RS\, \mathbf{s} = (2\pi/\lambda)\mathbf{s}$$

then
$$(\mathbf{G} + \mathbf{k})^2 = \mathbf{k}^2$$

or
$$2\mathbf{k}\cdot\mathbf{G} + G^2 = 0 \tag{5-14}$$

PROBLEM 5-25

(a) Show that Eq. (5-14) is the vector form of the Bragg law.

(b) Show that Eq. (5-14) defines a plane generated by a variable vector $\mathbf{k}$ which bisects, and is perpendicular to, a constant vector $\mathbf{G}$.

Since OS is a vector connecting the origin of reciprocal space with any lattice point, this problem tells us that an x-ray beam whose propagation vector $\mathbf{k}$ has the proper magnitude $k = 2\pi/\lambda$ and direction to fall on a perpendicular bisector RT of a line like OS will suffer Bragg reflection. To consider this more closely, Fig. 5-31 is derived from the body-centered reciprocal lattice of the face-centered cubic lattice. Between the origin and the nearest points we have drawn all the planes which are perpendicular bisectors of the connecting lines. The region bounded by the bisecting planes, which looks like a cube with the corners cut off, is called the *first Brillouin zone* of the reciprocal lattice. Because the bounding planes which determine the Brillouin zone are defined in terms of the propagation vector via Eq. (5-14), reciprocal space is also referred to as *k-space.* Turning back to Fig. 1-21, we can now realize that we have already encountered the idea of a one-dimensional Brillouin zone. It was shown that the forbidden energy bands occurred at values of k, the propagation constant of the electron-wave, such that

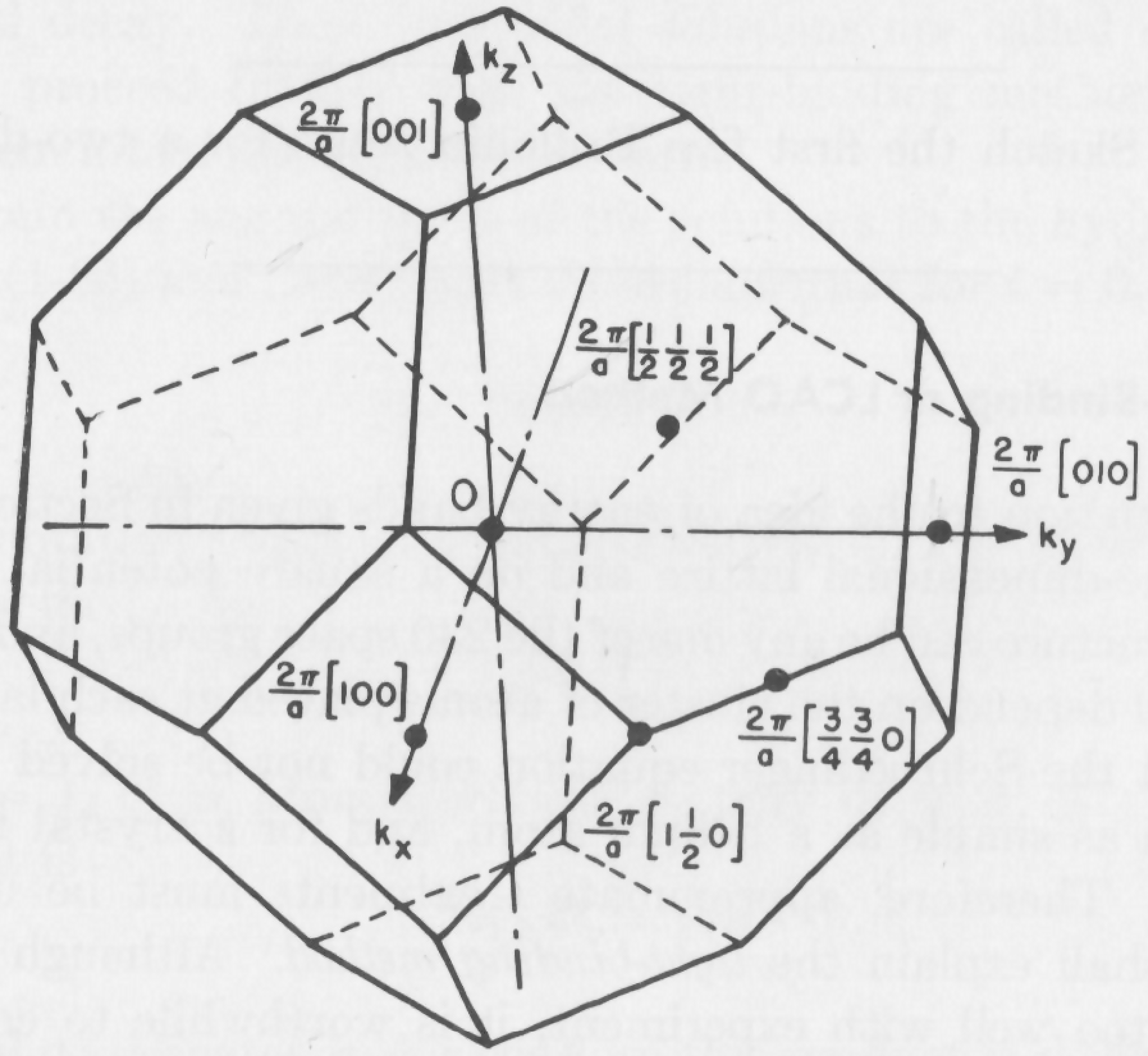

k_z
$\frac{2\pi}{a}[001]$
$\frac{2\pi}{a}[\frac{1}{2}\frac{1}{2}\frac{1}{2}]$
$\frac{2\pi}{a}[010]$
O
k_y
$\frac{2\pi}{a}[100]$
$\frac{2\pi}{a}[\frac{3}{4}\frac{3}{4}0]$
$\frac{2\pi}{a}[1\frac{1}{2}0]$
k_x

FIG. 5-31. The first Brillouin Zone for silicon and germanium. (From F. Herman, Proc. IRE **43,** 1703, 1955, with permission.)

$$k = n\pi/a \qquad (n = 0, 1, 2, 3) \tag{5-15}$$

PROBLEM 5-26

(a) Show that Eq. (5-15) is a form of the Bragg law.

(b) Show that the first Brillouin zone in a one-dimensional lattice lies in the region

$$-\pi/a \leqq k \leqq \pi/a$$

(c) Verify Fig. 5-31.

We can define Brillouin zones beyond the first by drawing perpendicular bisectors from the origin to the second and successive nearest neighbors. However, in dealing with energy bands, Fig. 3-1 shows that the periodic nature of crystals makes it possible to consider all the bands to be lying in the first zone. Such a zone is then known as a *reduced zone.*

PROBLEM 5-27 Sketch the first five Brillouin zones for a two-dimensional square lattice.

5-7 The Tight-Binding or LCAO Method

The introduction to the idea of energy bands given in Section 1-13 was based on a one-dimensional lattice and on a square potential. In a real crystal, the structure can be any one of the 230 space groups, and the potential energy will depend on the cluster of atoms placed at each lattice point. We know that the Schroedinger equation could not be solved exactly for anything even as simple as a helium atom, and for a crystal it is an impossible task. Therefore, approximate treatments must be used; as an example, we shall explain the *tight-binding method.* Although the results do not agree too well with experiment, it is worthwhile to consider this method as a simple introduction to band theory in real crystals.

Assume that we are dealing with a lattice whose atoms are so far apart that the behavior of an electron in the vicinity of any one of them is influenced only slightly by the other atoms (thus explaining the name). We then have the problem of solving the three-dimensional Schroedinger equation

$$-\frac{\hbar^2}{2m}\nabla^2\psi(\mathbf{r}) + V(\mathbf{r})\psi - E\psi = 0 \tag{1-79}$$

for this particular form of $V(\mathbf{r})$. We already know a few things about the nature of the solutions when $V(\mathbf{r})$ is periodic; we saw in Section 1-13 that the one-dimensional solution may have the form

$$\psi(x) = e^{ikx}u(x) \tag{1-102}$$

and in Section 3-1 it was shown that k is restricted to a discrete set of values. A distinct solution will exist for each value of $\mathbf{k}$. We denote this by writing $\psi_{\mathbf{k}}$.

Another property of the solution comes from the discussion of multielectron atoms in Section 1-12. It was pointed out there that the solution to the Schroedinger equation for all atoms beyond hydrogen is such a difficult task that to make even a qualitative estimate of the solution, we must make the approximation that the potential energy $V(\mathbf{r})$ is spherically symmetric. Doing this, however, helps a great deal, because the angular dependence of the solutions is the same as that for hydrogen, and they look like Fig. 1-17 if we ignore the shading. We cannot say anything about the dependence on r, but we can guess that there is some kind of exponential decay. These individual solutions are called *atomic orbitals.* Before we proceed further with the tight-binding method, we shall rearrange them into a more symmetric form.

We obtain the angular parts of the solutions to the hydrogen equation from Eqs. (1-95) and (1-96), and these show that for $l = 0$, the product is

$$\Theta_{0,0}\,\Phi_0 = 1$$

Using the notation s to denote the product $\Theta\Phi$ for $l = 0$ (*see* Section 1-12), we shall define

$$s = 1 \tag{5-16}$$

For the $l = 1$, or p, atomic orbitals m_l may be 0 or ± 1; when $m_l = 0$, the product is

$$\Theta_{1,0}\,\Phi_0 = \cos\theta$$

In spherical coordinates, $z = r\cos\theta$, so this product is designated as

$$p_z = \cos\theta \tag{5-17}$$

For $m_l = \pm 1$, it is convenient to take the solution as the sum of $e^{i\varphi}$ and $e^{-i\varphi}$, which are obtained from Eq. (1-95). We then have

$$p_x = \sin\theta(e^{i\varphi} + e^{-i\varphi}) = \sin\theta\cos\varphi \tag{5-18}$$

where any constants have been ignored.

PROBLEM 5-28

(a) Verify that Eq. (5-18) is a solution to the Schroedinger equation for hydrogen.

(b) Explain the notation p_x.

Similarly, we can also use the combination

$$p_y = \sin\theta(e^{i\varphi} - e^{-i\varphi}) = \sin\theta\sin\varphi \tag{5-19}$$

Going on to the d-functions, $l = 2$, we define in the same way

$$d_{z^2} = (3\cos^2\theta - 1) = \frac{3z^2}{r^2} - 1 \tag{5-20}$$

$$d_{xz} = \sin\theta\cos\theta\cos\varphi \tag{5-21}$$

$$d_{yz} = \sin\theta\cos\theta\sin\varphi \tag{5-22}$$

$$d_{x^2-y^2} = \sin^2\theta\cos 2\varphi \tag{5-23}$$

$$d_{xy} = \sin^2\theta\sin 2\varphi \tag{5-24}$$

PROBLEM 5-29 Identify the quantum numbers m_l and l for each of the d-functions and justify the subscripts for the last four.

The angular parts of the atomic orbitals are illustrated in Fig. 5-32, and it is seen that they are more symmetric than those of Fig. 1-17. For example, p_x, p_y, and p_z are identical except for orientation, and all are symmetric about the associated axis. The function d_{z^2} also has axial symmetry plus mirror symmetry in the XOY plane, and the other four d-functions are identical except for orientation.

Let us now consider an atom with position-vector $\mathbf{R}_j$. An electron with position-vector $\mathbf{r}$ will have a coordinate $(\mathbf{r} - \mathbf{R}_j)$ with respect to this atom, and we will denote the atomic orbitals for this electron by $\varphi(\mathbf{r} - \mathbf{R}_j)$. Now let us try a solution

$$\psi_{\mathbf{k}} = \sum_j e^{i\mathbf{k}\cdot\mathbf{R}_j}\,\varphi(\mathbf{r} - \mathbf{R}_j) \tag{5-25}$$

where the sum is taken over all the atoms in the crystal. The form of Eq. (5-25) leads to the alternate name: *linear combination of atomic orbitals* (*LCAO*); it is also called the *Bloch* method.

We would now like to show that Eq. (5-25) satisfies one of the properties for a solution in a periodic lattice. Let $\mathbf{r}_a$ denote the position-vector of a lattice-point (which is not necessarily the same as an atomic position $\mathbf{R}_j$). Substituting $\mathbf{r} + \mathbf{r}_a$ for $\mathbf{r}$ in Eq. (5-25) and multiplying the right-hand side by $e^{i\mathbf{k}\cdot\mathbf{r}_a}e^{-i\mathbf{k}\cdot\mathbf{r}_a}$ gives

$$\psi_{\mathbf{k}}(\mathbf{r} + \mathbf{r}_a) = e^{i\mathbf{k}\cdot\mathbf{r}_a}\sum_j e^{i\mathbf{k}\cdot(\mathbf{R}_j - \mathbf{r}_a)}\,\varphi[\mathbf{r} - (\mathbf{R}_j - \mathbf{r}_a)]$$

or

$$\psi(\mathbf{r} + \mathbf{r}_a) = e^{i\mathbf{k}\cdot\mathbf{r}_a}\,\psi(\mathbf{r}) \tag{5-26}$$

and this is seen to be the three-dimensional form of Eq. (1-102).

PROBLEM 5-30 Give the reasoning that leads to Eq. (5-26).

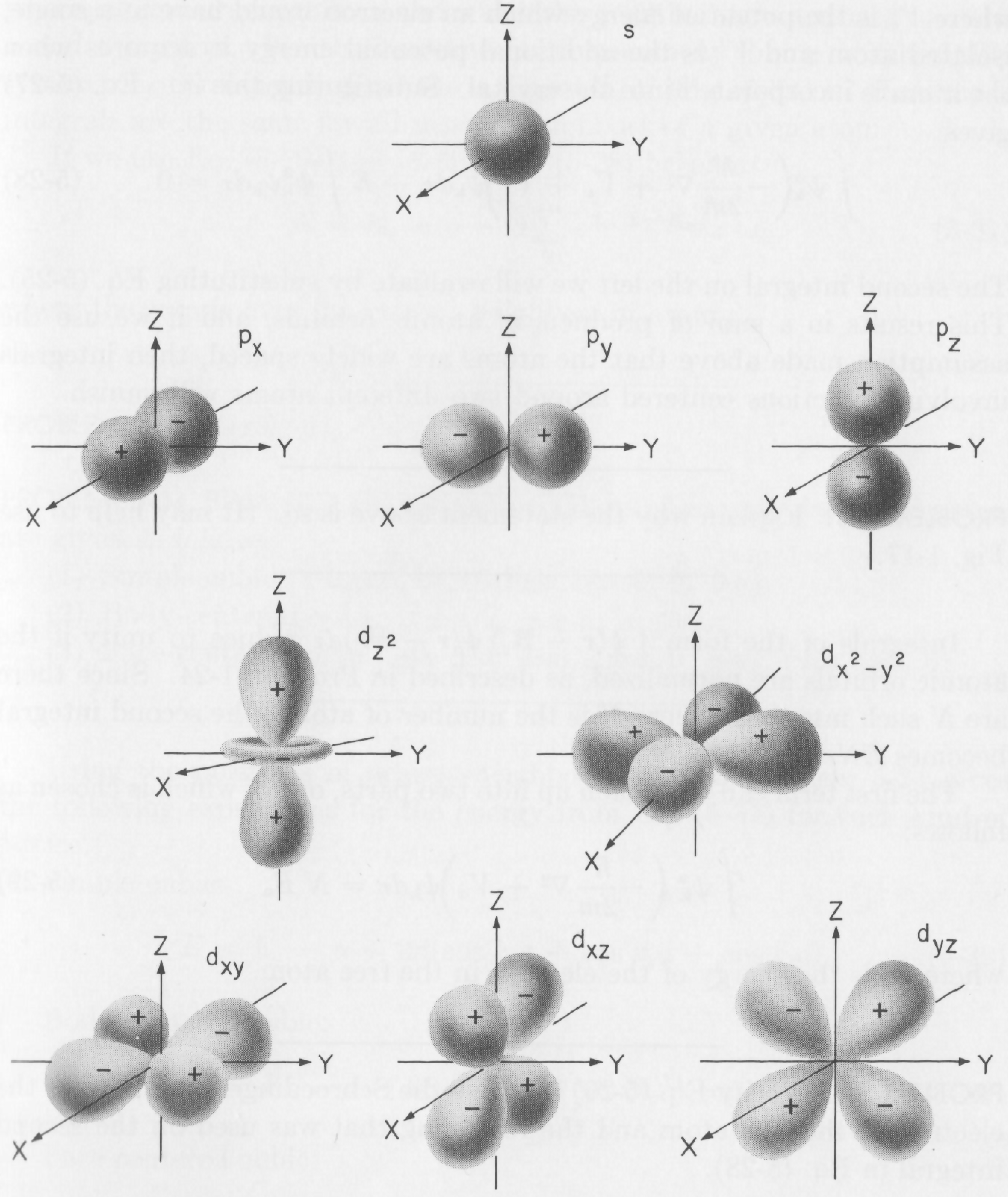


FIG. 5-32. The first nine atomic orbitals. (From H.B. Callen, *J. Chem. Phys.* **22**, 518, 1954.)

Now let us multiply the Schroedinger equation [Eq. (1-79)] for $\psi_{\mathbf{k}}$ by $\psi_{\mathbf{k}}^*$ and then integrate over all of space.

$$\int \psi_{\mathbf{k}}^*\left(\frac{-\hbar^2}{2m}\nabla^2 + V\right)\psi_{\mathbf{k}}\,d\tau + E\int \psi_{\mathbf{k}}^*\psi_{\mathbf{k}}\,d\tau = 0 \tag{5-27}$$

where the volume element is denoted $d\tau$ to avoid confusion with the symbol V for potential. Let us break V up into two terms.

$$V(\mathbf{r}) = V_a(\mathbf{r} - \mathbf{R}_j) + V'(\mathbf{r} - \mathbf{R}_j)$$

where V_a is the potential energy which an electron would have in a single, isolated atom and V' is the additional potential energy it acquires when the atom is incorporated into the crystal. Substituting this into Eq. (5-27) gives

$$\int \psi_{\mathbf{k}}^*\left(-\frac{\hbar^2}{2m}\nabla^2 + V_a + V'\right)\psi_{\mathbf{k}}d\tau - E\int \psi_{\mathbf{k}}^*\psi_{\mathbf{k}}d\tau = 0 \qquad (5\text{-}28)$$

The second integral on the left we will evaluate by substituting Eq. (5-25). This results in a sum of products of atomic orbitals, and if we use the assumption made above that the atoms are widely spaced, then integrals involving functions centered around two different atoms will vanish.

PROBLEM 5-31 Explain why the statement above is so. (It may help to use Fig. 1-17.)

Integrals of the form $\int \psi(\mathbf{r} - \mathbf{R}_j)\,\psi(\mathbf{r} - \mathbf{R}_j)d\tau$ reduce to unity if the atomic orbitals are normalized, as described in Problem 1-24. Since there are N such integrals, where N is the number of atoms, the second integral becomes EN.

The first term can be broken up into two parts, one of which is chosen as follows:

$$\int \psi_{\mathbf{k}}^*\left(-\frac{\hbar^2}{2m}\nabla^2 + V_a\right)\psi_{\mathbf{k}}d\tau = N\,E_a \qquad (5\text{-}29)$$

where E_a is the energy of the electron in the free atom.

PROBLEM 5-32 Verify Eq. (5-29) by using the Schroedinger equation for the electron in the free atom and the reasoning that was used on the second integral in Eq. (5-28).

This leaves only the integral involving V' to be considered; here we shall neglect all terms in the summations except those involving immediate neighbors, an inherent feature of the tight-binding assumption. Define

$$\begin{aligned} \alpha &= -N\int \varphi^*(\mathbf{r} - \mathbf{R}_j)V'\,\varphi(\mathbf{r} - \mathbf{R}_j)d\tau \\ \beta &= -N\int \varphi^*(\mathbf{r} - \mathbf{R}_m)V'\,\varphi(\mathbf{r} - \mathbf{R}_j)d\tau \end{aligned} \qquad (5\text{-}30)$$

where j and m are neighbors, and where the negative signs are used to make α and β positive. We shall assume for the moment that φ is spherically symmetric, corresponding to the s function of Fig. 5-32, so that the β integrals are the same for all nearest neighbors of a given atom.

If we use Eq. (5-29) and (5-30), Eq. (5-28) becomes

$$E = E_a - \alpha - \beta\sum_m e^{i\mathbf{k}\cdot(\mathbf{R}_j - \mathbf{R}_m)} \qquad (5\text{-}31)$$

where the sum is over the nearest neighbors of atom j.

PROBLEM 5-33 Verify Eq. (5-31).

PROBLEM 5-34 Show that the nearest neighbors of an atom in cubic lattices are given as follows:

(1) Simple cubic: $(\pm a, 0, 0)$, $(0, \pm a, 0)$, $(0, 0, \pm a)$.
(2) Body-centered cubic: $(\pm a/2, \pm a/2, \pm a/2)$.
(3) Face-centered cubic: $(0, \pm a, \pm a)$, $(\pm a, 0, \pm a)$, $(\pm a, \pm a, 0)$.

Using the positions of nearest-neighbors listed in Problem 5-34 gives the following expressions for the energy from Eq. (5-31) for each kind of lattice.

Simple cubic:

$$E = E_a - \alpha - 2\beta(\cos k_x a + \cos k_y a + \cos k_z a) \qquad (5\text{-}32)$$

Body-centered cubic:

$$E = E_a - \alpha - 8\beta\cos(k_x a)\cos(k_y a)\cos(k_z a) \qquad (5\text{-}33)$$

Face-centered cubic:

$$\begin{aligned} E = E_a - \alpha - 4\beta[&\cos(k_x a)\cos(k_y a) + \cos(k_x a)\cos(k_z a) \\ &+ \cos(k_y a)\cos(k_z a)] \end{aligned} \qquad (5\text{-}34)$$

These three equations express E as a function of k_x, k_y, and k_z. To visualize these relations, consider a two-dimensional square lattice, so that $k_z = 0$ in Eq. (5-32). We then wish to plot

$$E - E_a + \alpha = -2\beta(\cos k_x a + \cos k_y a)$$

where E_a, α, and β are unknown constants. Since these constants do not affect the form of the relation between E and $\mathbf{k}$, we can replace E by a *reduced energy* E_r, defined as

$$E_r = \frac{E - E_a + \alpha}{2\beta} = -(\cos k_x x + k_y y) \tag{5-35}$$

Curves of k_x vs k_y for various values of E_r are shown in Fig. 5-33.

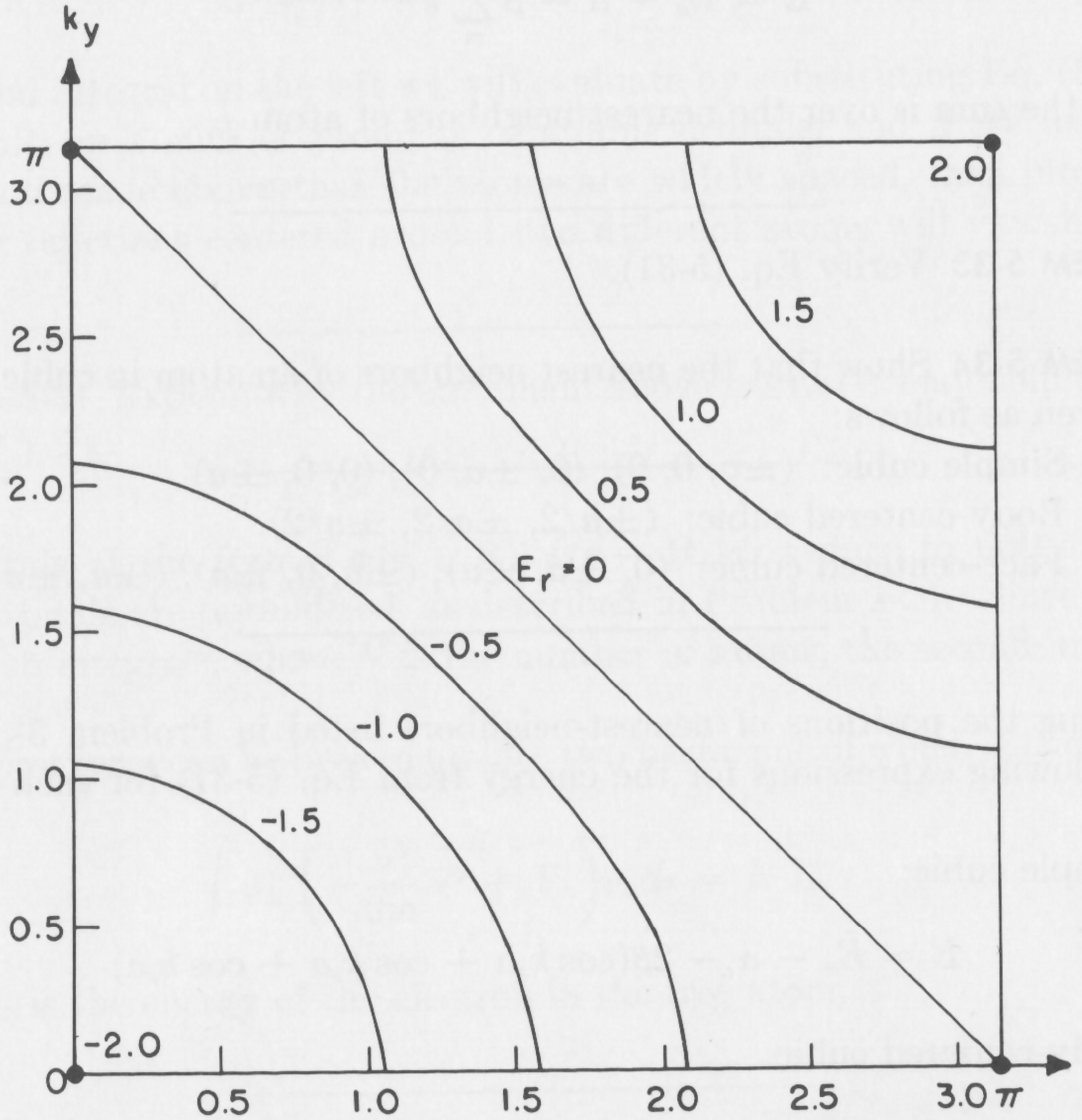


FIG. 5-33. The energy contours obtained when the tight-binding method is applied to a square lattice.

PROBLEM 5-35

(a) What is the significance of the limits on k_x and k_y shown in the figure?

(b) Show by the use of a Taylor series for cosines (*see* Appendix A-1) that the curves for k_x and k_y close to the origin are circles.

(c) Show that the same is true for k_x and k_y close to the upper right-hand corner.

(d) Show that the curves for $E_r = 0$ are straight lines through k_x, $k_y = \pm\pi/a$.

Next, we shall consider the tight-binding approximation in the simple cubic lattice when the functions $\varphi(\mathbf{r} - \mathbf{R}_j)$ are taken as one of the p functions of Fig. 5-32. The integrals of Eq. (5-30) are different for p functions from those for s functions, and further, the integrals are not the same for all nearest neighbors because of the directional properties of p_x, p_y, and p_z. For convenience, consider just the p_x function, and denote the integral corresponding to α by α_p.

PROBLEM 5-36

(a) Consider a function p_x located at the origin, and six more located at each of six nearest neighbors in a simple cubic lattice. Show that there are two kinds of integrals of the form of β in Eq. (5-30), denoted β_{p1} and β_{p2}.

(b) Show that

$$E = E_{ap} - \alpha_p - 2\beta_{p1}\cos k_x a - 2\beta_{p2}(\cos k_y a + \cos k_z a) \tag{5-36}$$

where

$$\alpha_p = -N\int \varphi^*(0)V'\varphi(0)\,d\tau$$

$$\beta_{p1} = -N\int \varphi^*(a, 0, 0)V'\varphi(0)\,d\tau$$

$$\beta_{p2} = -N\int \varphi^*(0, a, 0)V'\varphi(0)\,d\tau$$

and $\varphi(0)$ is a function located at the origin, $\varphi(a, 0, 0)$ is located at $x = a$, and $\varphi(0, a, 0)$ is located at $y = a$ (moving this function to $y = -a$ or to $z = \pm a$ gives the same integral).

We can determine the signs of the integrals in Eq. (5-36) by a method of Mott and Jones (1936).[13] The integral involving a p_x function at the origin and one at $x = a$ is

$$\beta_{p1} = \int p_x^*(a)V'p_x(0)d\tau$$

and at a point midway between the two atoms, the two p_x functions have the maximum *overlap*, since the lobes of both functions lie along the x-axis. The overlapping lobes have the opposite sign, and since V' is negative (Fig. 1-19 shows the approximate behavior of the potential of the electron), the product of the three terms in the integral is positive. On the other hand, for integrals involving a p_x function at y or $z = \pm a$, the integrand midway between two atoms is negative. Further, the overlap is very small and the value of the integral is much smaller.

PROBLEM 5-37

(a) Show that α_p is positive.

(b) Neglecting β_{p2}, show that E decreases with k along the k_x-axis, giving the upper curve of Fig. 5-34.

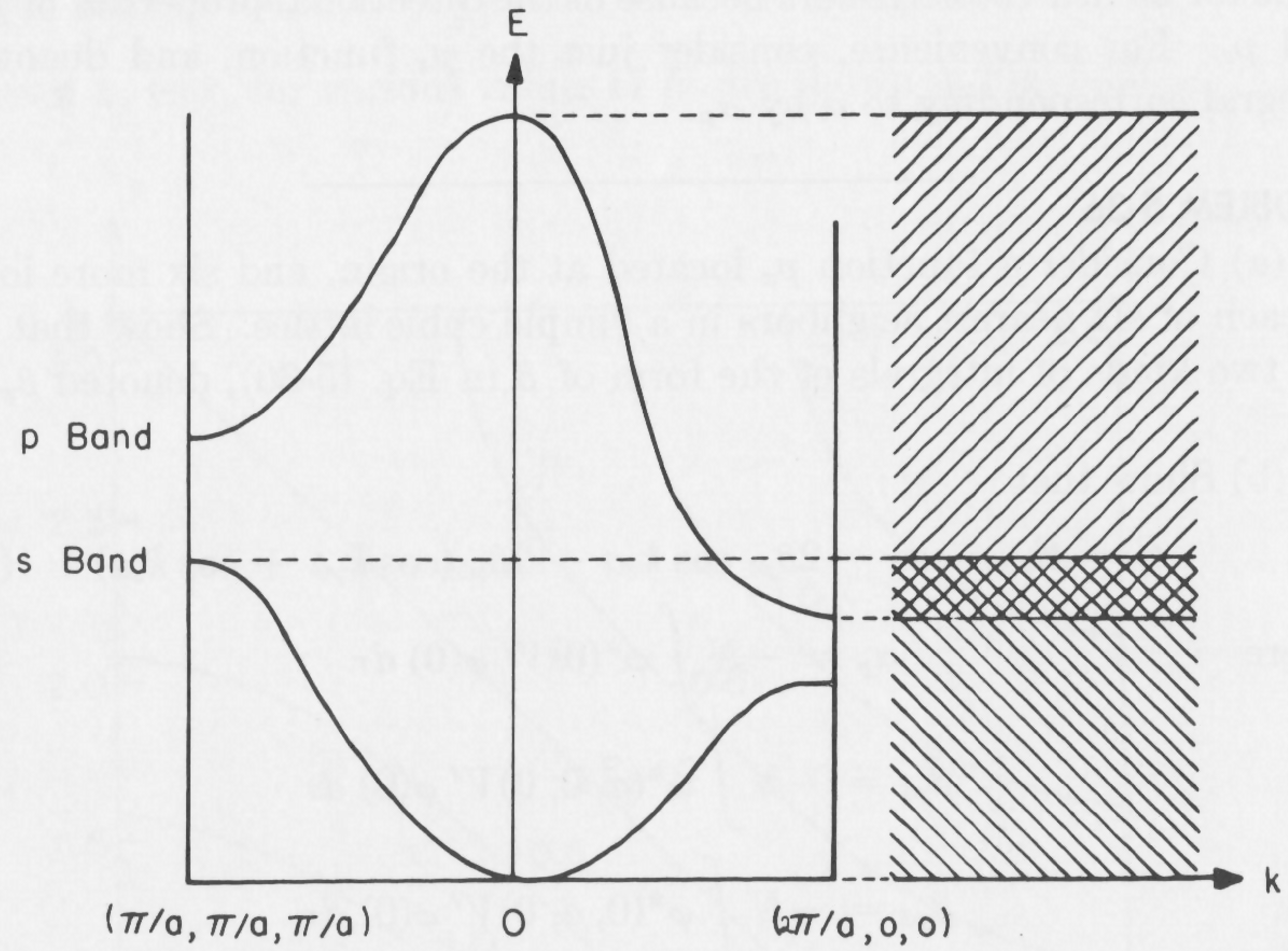


FIG. 5-34. Overlap obtained by use of p-functions in the tight-binding method. (From J. M. Ziman, *Electrons and Photons*, Oxford Univ. Press, London, 1960, with permission.)

In Fig. 5-34, we have plotted the curve for p-functions in order to provide an energy gap between the two allowed bands. A feature of energy bands in two and three dimensions which is not found in one dimension is the possibility of overlapping bands. The left-hand side of Fig. 5-34 shows the E vs k curves along the cube diagonal of the Brillouin zone.

PROBLEM 5-38 Verify that the two sides of the curve for s functions in the figure are drawn to the same scale.

We see from Fig. 5-34 that energies which are forbidden to an electron whose propagation vector lies along the x-axis of k-space are allowed for an electron with a vector along the cube diagonal. Hence, all energies on the diagram are available and the material is a good conductor. We mentioned in Section 3-1 that metals are characterized by partially filled bands, and we see now we can also get metallic conduction with overlapping bands.

5-8 The Band Structure of Tellurium

As an example of the application of the tight-binding method to a real crystal, let us consider the determination of the band structure in tellurium made by Reitz (1957).[14] Table 1-5 shows that the six valence electrons of tellurium (element 52) are designated $5s^2p^4$; Reitz, therefore, took the p functions as being those involved in the valence and conduction bands.

For convenience, denote the functions $\varphi(\mathbf{r} - \mathbf{R}_j)$ by $\varphi_1 = p_x$, $\varphi_2 = p_y$, and $\varphi_3 = p_z$. Then there are three expressions of the form of Eq. (5-25).

$$\psi_{n\mathbf{k}} = \sum_j e^{i\mathbf{k}\cdot\mathbf{R}_j}\,\varphi_n(\mathbf{r} - \mathbf{R}_j) \qquad (n = 1, 2, 3)$$

and we shall try as a solution of the Schroedinger equation the sum of these three individual solutions, written as

$$\psi_{\mathbf{k}}(\mathbf{r}) = \sum_{j,n,s} B_{ns}e^{i\mathbf{k}\cdot\mathbf{R}_{js}}\,\varphi_n(\mathbf{r} - \mathbf{R}_{js}) \tag{5-37}$$

where the B_{ns} are constants to be determined. The subscript s is introduced into Eq. (5-37) because tellurium has three atoms per unit cell, as we saw in Section 5-5, and we sum also over these non-equivalent lattice sites.

Again writing the potential term V in the Schroedinger equation as $V = V_a + V'$ and substituting Eq. (5-37) into it gives

$$\sum_{j,n,s} B_{ns}\,e^{i\mathbf{k}\cdot\mathbf{R}_{js}}\left(-\frac{\hbar^2}{2m}\nabla^2 + V_a + V' - E\right)\varphi_n(\mathbf{r} - \mathbf{R}_{js}) = 0 \tag{5-38}$$

Multiplying this by each one of ψ_1^*, ψ_2^*, and ψ_3^* in turn and integrating gives nine equations in the nine unknowns B_{ns}. These equations are *homogeneous* — there are no terms not involving the unknowns — and the condition that such a set of equations have consistent solutions is that the determinant of the coefficients of the B_{ns} vanish. Using this criterion, we obtain what is known as the *secular equation.* To apply this to tellurium, let us first make the simplifying assumption that the angles between the lines connecting adjacent atoms in a chain are 90° instead of the known value of 102.6°. Label the three atoms in the unit cell as A, B, and C (Fig. 5-35) and let $\mathbf{R}_\lambda$, $\mathbf{R}_\mu$, and $\mathbf{R}_\nu$ denote the three nearest-neighbor distances, so that the lattice constant along the trigonal axis can be expressed as the vector

$$\mathbf{c} = -\mathbf{R}_\lambda + \mathbf{R}_\mu + \mathbf{R}_\nu$$

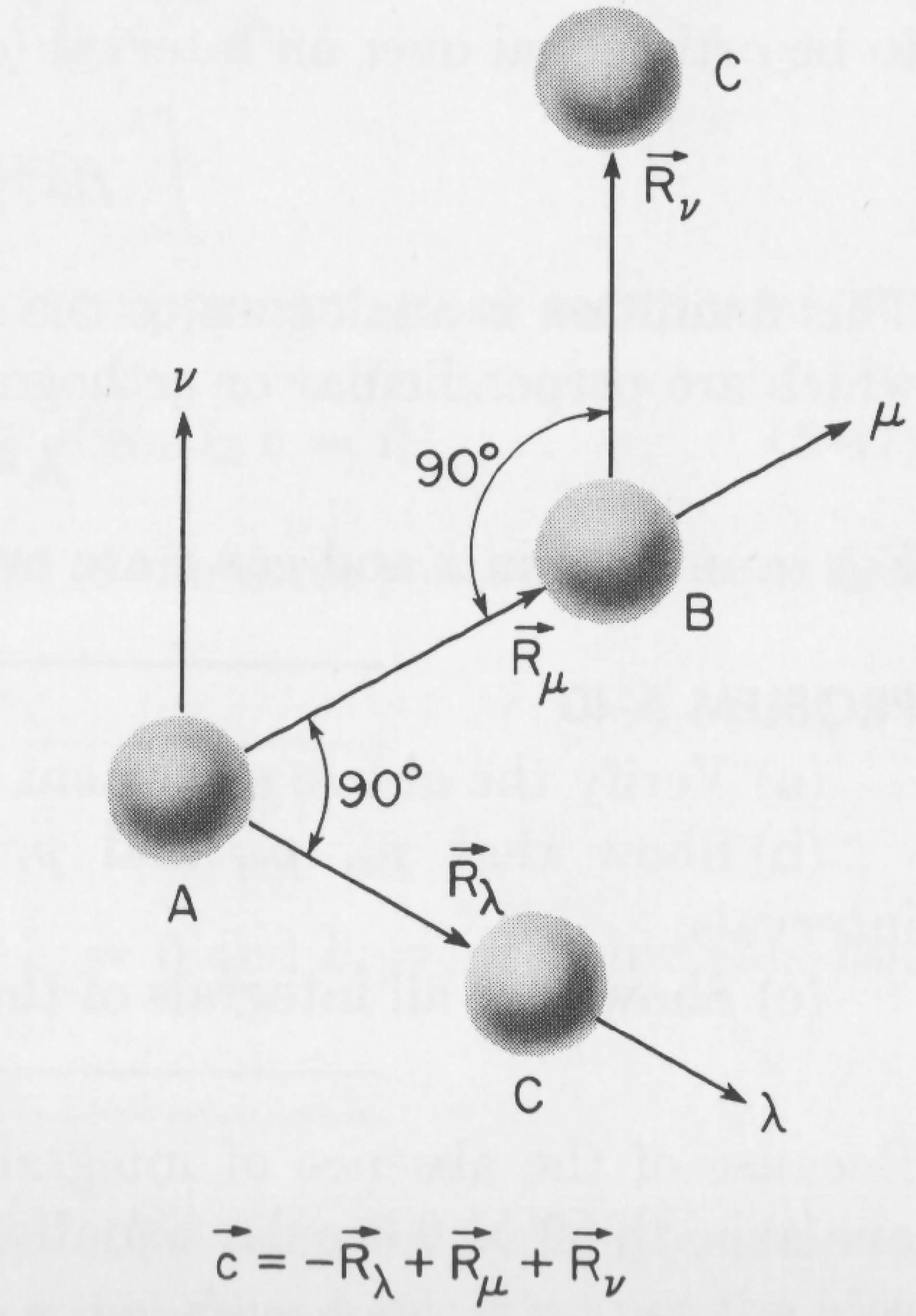


FIG. 5-35. The approximation of 90° bonds for tellurium chains. (From J. R. Reitz, Phys. Rev. **105**, 1233, 1957, with permission.)

We then orient the p functions so that they lie along the λ, μ, and ν axes. Because there are so many possible integrals of the form of α_p and β_p of the last section, we use a special symbol for them, as follows:

$$(\varphi_n^*|V'|\varphi_m) = N\int \varphi_n^* V'\,\varphi_m\,d\tau \tag{5-39}$$

If only nearest-neighbor integrals are permitted, and if the location of a function is denoted by a subscript A, B, or C, then the integrals involving p_λ functions are

$$(p_{\lambda A}|V'|p_{\lambda A}) = (p_{\lambda C}|V'|p_{\lambda C}) \tag{5-40}$$

$$(p_{\lambda B}|V'|p_{\lambda B}) \tag{5-41}$$

$$(p_{\lambda A}|V'|p_{\lambda B}) = (p_{\lambda B}|V'|p_{\lambda A}) = (p_{\lambda B}|V'|p_{\lambda C}) = (p_{\lambda C}|V'|p_{\lambda B}) \tag{5-42}$$

$$(p_{\lambda C}|V'|p_{\lambda A}) = (p_{\lambda A}|V'|p_{\lambda C}) \tag{5-43}$$

PROBLEM 5-39 Why are the integrals in Eqs. (5-40) and (5-41) different? Why are the four integrals in Eq. (5-42) the same, but different from those in Eq. (5-43)?

Integrals of the form $(p_{\lambda A}|V'|p_{\mu A})$ are zero because of a property of the p functions known as *orthogonality*. Two functions $f(x)$ and (gx) are said to be orthogonal over an interval (a, b) if

$$\int_a^b f(x)\,g(x)\,dx = 0 \tag{5-44}$$

This definition is analogous to the scalar product of two vectors **A** and **B**, which are perpendicular or orthogonal if

$$\mathbf{A}\cdot\mathbf{B} = 0$$

For example, $\sin x$ and $\cos x$ are orthogonal over the interval 0 to 2π.

PROBLEM 5-40

(a) Verify the above statement.

(b) Show that p_x, p_y, and p_z are orthogonal over the appropriate intervals.

(c) Show that all integrals of the form $(p_{\lambda A}|V'|p_{\mu A})$ are zero.

Because of the absence of integrals involving functions along more than one axis, the 9×9 secular equation has the form

$$\begin{matrix} \lambda & 0 & 0 \\ 0 & \mu & 0 \\ 0 & 0 & \nu \end{matrix} = 0 \tag{5-45}$$

where λ denotes a 3×3 block involving p_λ functions only, μ and ν are of the same form for the two other p functions, and the other six blocks are composed entirely of zeros.

PROBLEM 5-41 Show that each block in Eq. (5-45) can be equated separately to zero.

If we use Problem 5-41, the λ block becomes

$$\begin{vmatrix} (p_{\lambda A}|V'|p_{\lambda A}) + E_a - E & (p_{\lambda A}|V'|p_{\lambda B})e^{i\mathbf{k}\cdot\mathbf{R}_\mu} & (p_{\lambda A}|V'|p_{\lambda C})e^{i\mathbf{k}\cdot\mathbf{R}\lambda} \\ (p_{\lambda B}|V'|p_{\lambda A})e^{-i\lambda\cdot\mathbf{R}_\mu} & (p_{\lambda B}|V'|p_{\lambda B}) + E_a - E & (p_{\lambda B}|V'|p_{\lambda C})e^{i\mathbf{k}\cdot\mathbf{R}} \\ (p_{\lambda C}|V'|p_{\lambda A})e^{-i\mathbf{k}\cdot\mathbf{R}\lambda} & (p_{\lambda C}|V'|p_{\lambda B})e^{-i\mathbf{k}\cdot\mathbf{R}_\nu} & (p_{\lambda C}|V'|p_{\lambda C}) + E_a - E \end{vmatrix} = 0 \tag{5-46}$$

In order to solve this cubic equation for E, we make some further approximations. We estimate that the integrals of the form $(p_{\lambda A}|V'|p_{\lambda B})$ have a magnitude about one-third of those like $(p_{\lambda A}|V'|p_{\lambda C})$ and we assume that the integrals in Eqs. (5-40) and (5-41) are the same. If we introduce a relative energy E_R by

$$E_R = E - E_a - (p_{\lambda A}|V'|p_{\lambda A})$$

and the abbreviation

$$\sigma = (p_{\lambda A}|V'|p_{\lambda C})$$

the secular equation [Eq. (5-46)] with the approximations mentioned becomes

$$-E_R^3 + \tfrac{11}{9}\,E_R\sigma^2 + \tfrac{2}{9}\,\sigma^3\cos k_z\,c = 0 \tag{5-47}$$

The dotted curves of Fig. 5-36 show a plot of the three roots of this equation as a function of k_z.

PROBLEM 5-42

(a) Verify Eq. (5-47).

(b) Find the values of the energy at $k_z = 0$ and $k_z = \pi/c$ which are not shown for the dotted curves of Fig. 5-36.

The same set of dotted curves will also come from the μ and ν equations, and the bands are, therefore, said to be *triply degenerate*, if we use the idea of degeneracy defined in Section 3-1. If the 90-deg angles in Fig. 5-35 are now stretched to their true value of 102.6°, we introduce a number of additional integrals of the form $(p_{\lambda A}|V'|p_{\mu B})$ into the secular determinant,

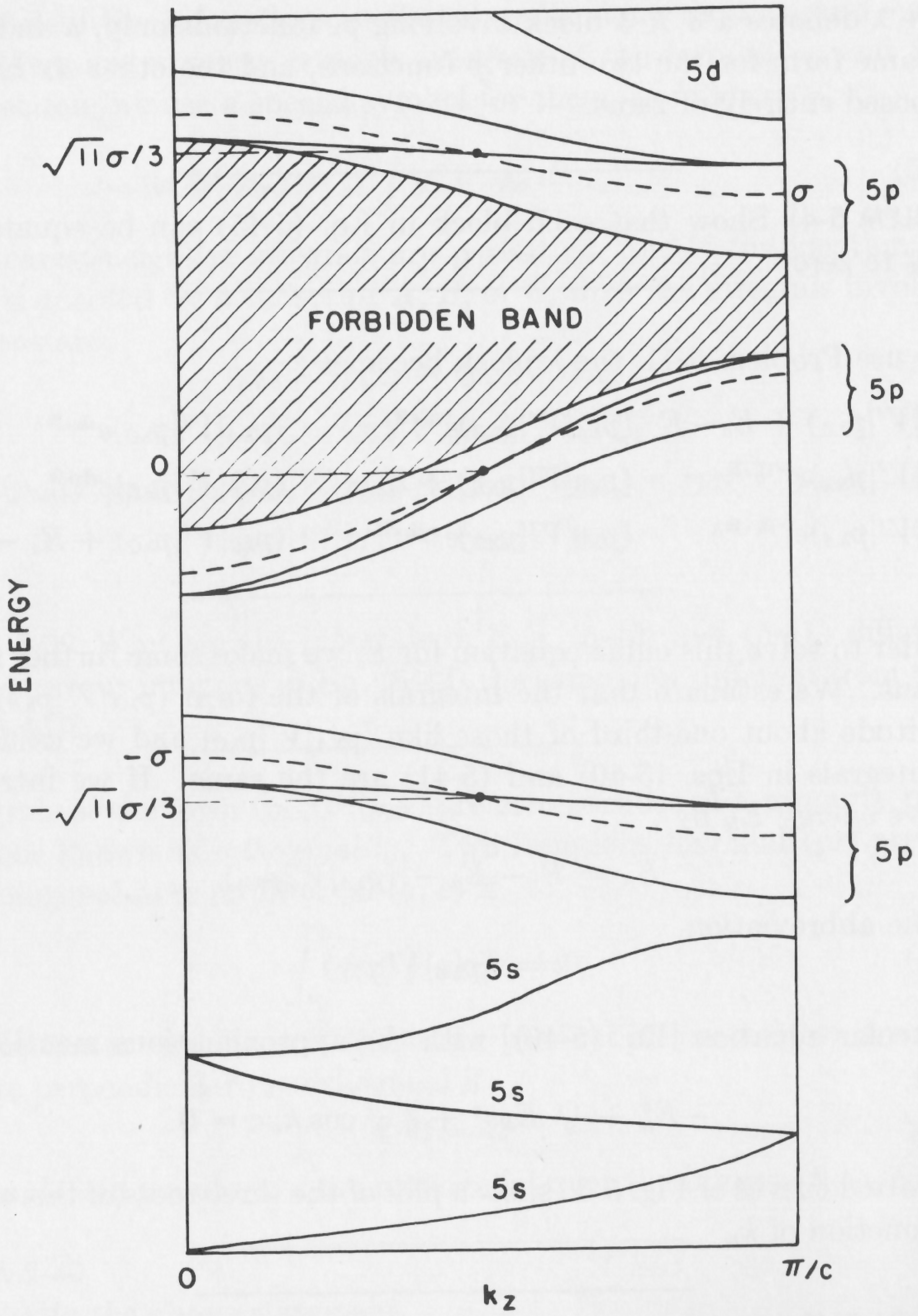


FIG. 5-36. The results of the tight-binding method applied to tellurium. (From Reitz, *op. cit.*)

and E_R can be obtained only when some additional approximations are made. The solutions of the secular equation are shown as the solid lines labelled $5p$ bands in Fig. 5-36. We see that each dotted curve is split up into three solid ones; the stretching of the interatomic angles is said to *remove the degeneracy*. Note, however, that two-fold degeneracy exists at $k = 0$ and π/c. We have added to this figure curves for the s band; these curves are schematic rather than being calculated, with the results of the previous section as a guide. Since tellurium has three atoms per unit cell and there is one s function for each atom, there are three s bands shown, and, similarly, there are nine p bands. The position of the forbidden band is determined from the fact that there are eighteen electrons in the unit cell; with two per band, considering spin, the top of the valence band is the ninth one. We shall not look into the significance of this energy diagram for tellurium any further at this point, but it will be considered in connection with the experimental properties treated in Chapter 7.

5-9 The Energy Bands of Silicon and Germanium

Although the tight-binding method described in the previous section has been applied to column IV semiconductors, the most accurate calculations of the band structure of diamond-type crystals have been those of Herman (1955)[15], who has used the method of *orthogonalized plane waves* (*OPW*). The name comes from the type of solution tried in the Schroedinger equation. For the LCAO method, we used

$$\psi_{\mathbf{k}} = \sum_j e^{i\mathbf{k}\cdot\mathbf{R}_j}\,\varphi(\mathbf{r} - \mathbf{R}_j) \tag{5-25}$$

where the $\varphi(\mathbf{r} - \mathbf{R}_j)$ were taken as s or p functions. In the OPW method, these atomic orbitals are replaced by *plane-wave* functions $e^{i\mathbf{k}\cdot\mathbf{r}}$, where

$$\mathbf{h} = h_1\mathbf{b}_1 + h_2\mathbf{b}_2 + h_3\mathbf{b}_3 \qquad (h_1, h_2, h_3 = 1, 2, 3, \ldots)$$

PROBLEM 5-43 Show that $e^{i\mathbf{h}\cdot\mathbf{r}}$ represents a wave of constant phase in direct space.

A solution is then tried of the form

$$\psi_{\mathbf{k}} = \sum_{h_1,h_2,h_3} e^{i(\mathbf{k}+\mathbf{h})\cdot\mathbf{r}}$$

The condition that this solution satisfies the Schroedinger equation is a secular equation like Eq. (5-45). We have seen that by making some reasonable approximations, Eq. (5-45) reduced to Eq. (5-47), which involves k_z as the independent variable. For an accurate calculation of the E vs $\mathbf{k}$ curves in silicon and germanium, it is necessary to solve a number of secular equations, one for each type of symmetry of $\mathbf{k}$ in the Brillouin zone. Referring to Fig. 5-31, we see that $\mathbf{k}$ has symmetry properties in the Brillouin zone in the same way that the crystal lattices show symmetry in direct space. The most symmetrical $\mathbf{k}$ value is the center point (0, 0, 0), and the coordinate axes are lines of symmetry. These are the k_x-axis, which extends from (0, 0, 0) to $2/\pi a\ (1, 0, 0)$, and the other two equivalent axes, denoted k_y and k_z.

PROBLEM 5-44 Show that other symmetry axes in the Brillouin zone run from (0, 0, 0) to $2\pi/a\ (\frac{3}{4}, \frac{3}{4}, 0)$ and to $2\pi/a\ (\frac{1}{2}, \frac{1}{2}, \frac{1}{2})$.

To find E vs $\mathbf{k}$ along these various directions in k-space involves solving determinants of high order, and the problem can be considerably simplified by using the methods of a branch of algebra known as *group theory*. We shall not give any details here, since they are somewhat lengthy, but we refer the interested student to the introductory books by Mariot (1959)[16] and Jones (1960)[17] (which should be read in the order mentioned). A sufficient background to understand Herman's results, taken from his excellent review article[15] and presented in Fig. 5-37, is provided by the discussion of the tellurium or selenium band structure of the previous section.

Let us concentrate for the moment on Fig. 5-37b, the band structure of germanium. We see that the E vs $\mathbf{k}$ curves are plotted as a function of distance along the various symmetry axes in the Brillouin zone of Fig. 5-31, so that instead of having one set of curves, as in Fig. 5-36, we have four sets. The notation "double" means that the secular equation has two sets of identical roots, just as it had three sets for the dotted curves of Fig. 5-36. The lower germanium band picture, Fig. 5-37d, shows the removal of this degeneracy by *spin-orbit coupling*. Both the orbital and spinning motion of the electron have magnetic fields associated with them, and the interaction of these two forces involves the amount of energy represented by the depression of the third valence band.

One of the most significant features of these diagrams is the location of the lowest point in the conduction band and the highest point in the valence band, since the vertical separation of these two extrema is the thermal energy gap of the semiconductor. We see that for germanium, there are minima at several places in the conduction band, the lowest one being at the center of the hexagonal face in the Brillouin zone. It should be realized that the E vs $\mathbf{k}$ curves in this figure are only one-dimensional; three-dimensional constant-energy surface analogous to the two-dimensional contours of Fig. 5-33 will be quite complicated. However, some very important information about the nature of the constant-energy surfaces at energies corresponding to the bottom of the conduction band — the energy region which is occupied by the conduction electrons — can be obtained by using a Taylor series. Following Herman (1955)[15], let k_0 denote the value

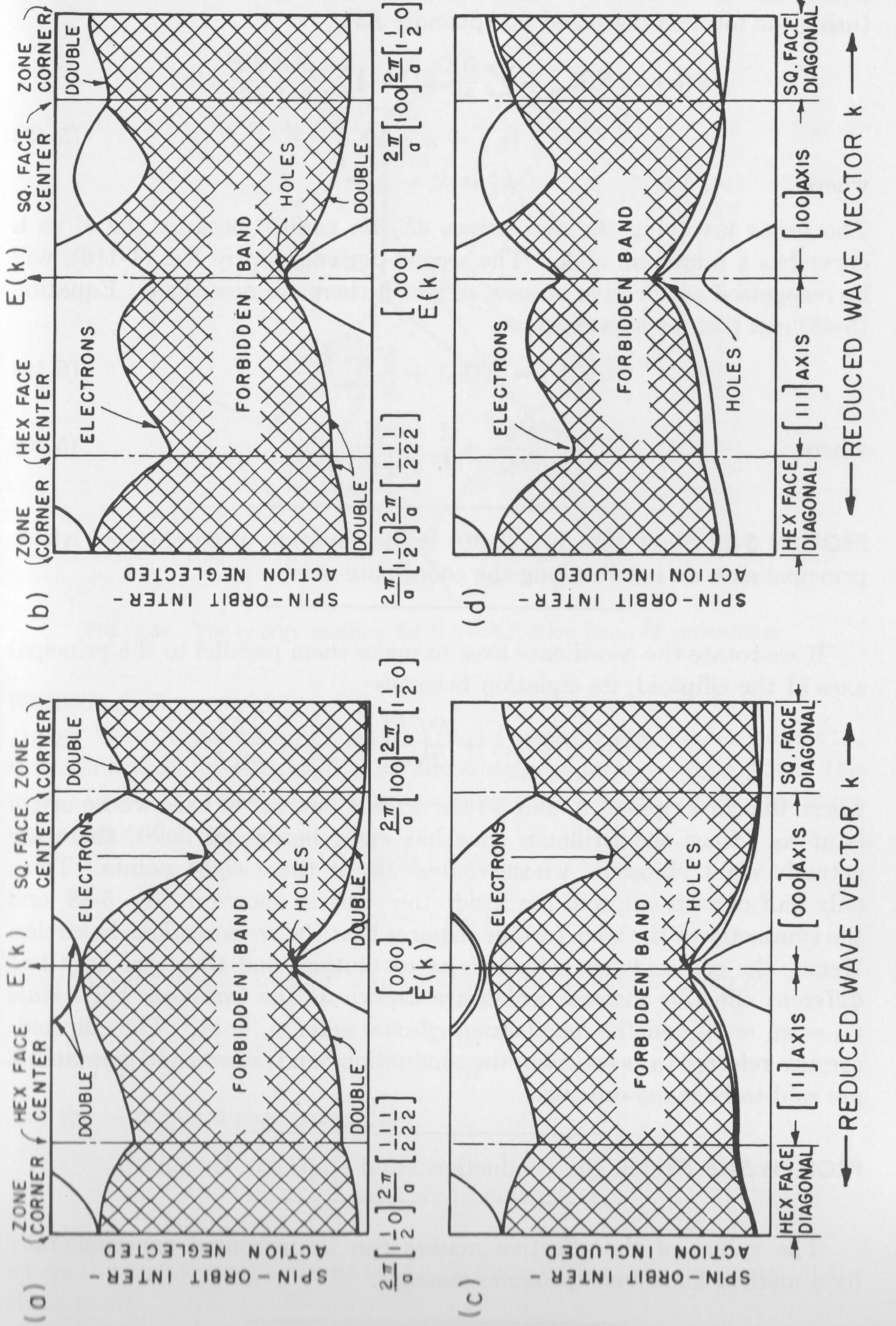


FIG. 5-37. The band structure of silicon and germanium. (From Herman, *op. cit.*)

of **k** at the center of the hexagonal face. Then, for values of **k** close to $\mathbf{k}_0$ (using the formula discussed in Appendix A1),

$$E(\mathbf{k}) = E(\mathbf{k}_0) + \sum_i \frac{\partial E}{\partial k_i} k'_i + \tfrac{1}{2} \sum_{ij} \frac{\partial^2 E}{\partial k_i^2} \frac{\partial^2 E}{\partial k_j^2} k'_i k'_j$$

$$(i, j = x, y, z) \qquad (5\text{-}48)$$

where

$$k'_i = k_i - k_0$$

The terms involving the derivatives $\partial E/\partial k_i$ vanish, because the E vs **k** curve has a minimum at $\mathbf{k}_0$. The second-derivatives, by Eq. (1-116), will be recognized as effective masses, of which there are now three. Equation (5-48) can then be expressed as

$$E(\mathbf{k}) = E(\mathbf{k}_0) + \frac{\hbar^2}{2} \sum_{i,j} \frac{k'_i k'_j}{m_{ij}} \qquad (5\text{-}49)$$

where

$$\frac{1}{m_{ij}} = \frac{1}{\hbar^2} \frac{\partial^2 E}{\partial k'_i \partial k'_j} \qquad (5\text{-}50)$$

PROBLEM 5-45 Show that Eq. (5-49) is the equation of an ellipsoid whose principal axes do not lie along the coordinate axes.

If we rotate the coordinate axes to make them parallel to the principal axes of the ellipsoid, its equation becomes

$$E(\mathbf{k}) = E(\mathbf{k}_0) + \frac{\hbar^2}{2}\left(\frac{k_1^2}{m_1^2} + \frac{k_2^2}{m_2^2} + \frac{k_3^2}{m_3^2}\right) \qquad (5\text{-}51)$$

where the subscripts 1, 2, and 3 refer to this new set of axes whose origin is at $\mathbf{k}_0$. Since the Brillouin zone has eight hexagonal faces, there are actually eight ellipsoids whose centers lie at these eight points. Thus, only half of each ellipsoid lies inside the zone, as shown in Fig. 5-38, and the symmetry of the zone further requires that the cross sections be circles. Hence, the ellipsoids are ellipsoids of revolution, and there are only two different effective masses, which are known as the *transverse effect mass* $m_t = m_1 = m_2$ and the *longitudinal effective mass* $m_l = m_3$. The ellipsoids are also referred to as *valleys* in the conduction band; silicon and germanium are said to be *many-valleyed*.

PROBLEM 5-46 Sketch the conduction band ellipsoids for silicon.

The values of the effective masses can be measured experimentally by a method known as *cyclotron resonance*.

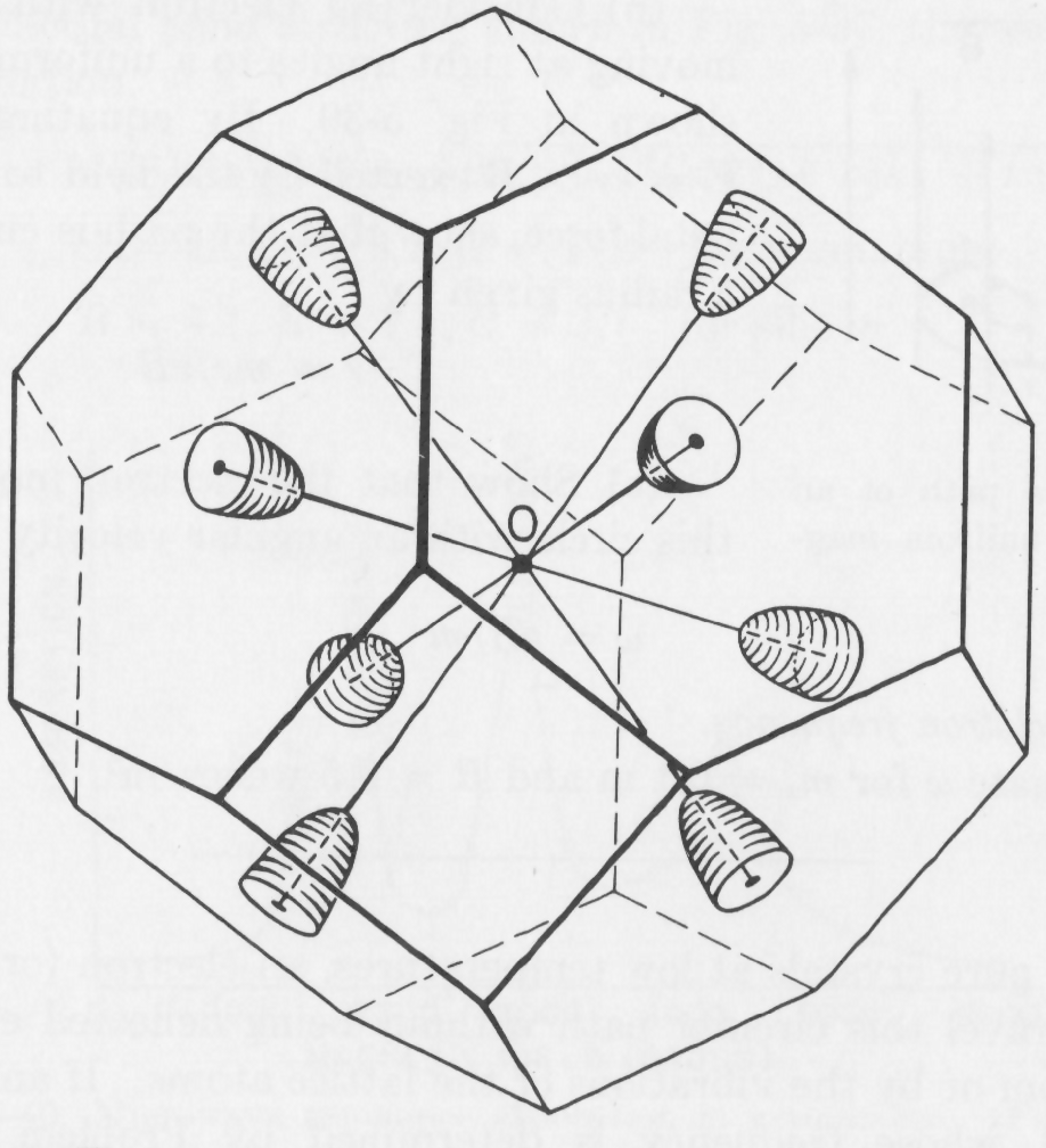


FIG. 5-38. The energy minima for the conduction band of germanium.

PROBLEM 5-47

(a) When a particle moves in a curved path, it takes a force directed towards the center of curvature — the *centripetal force* — to keep it in this path. Using the relations between rectangular and polar coordinates given in the appendix

$$x = r \cos \theta$$

$$y = r \sin \theta$$

show that for motion in a circle of radius r, the velocity is given by

$$v = r \frac{d\theta}{dt}$$

and the centripetal acceleration is

$$a_r = \frac{d^2x}{dt^2} \cos \theta + \frac{d^2y}{dt^2} \sin \theta = -\frac{v^2}{r}$$

where the negative sign means that a_r is directed towards the center of the circle.

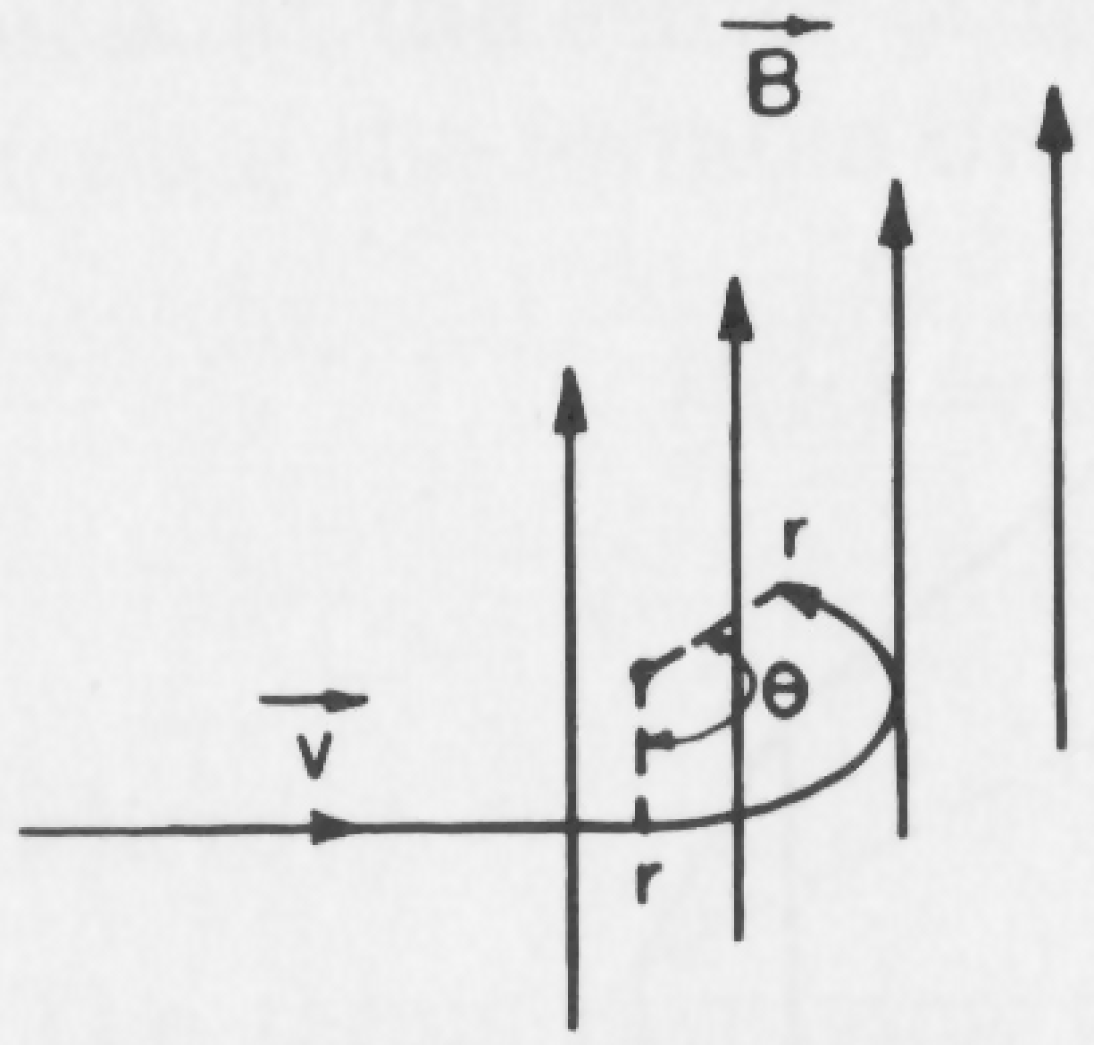


FIG. 5-39. The path of an electron in a unifrom magnetic field.

(b) Consider an electron with velocity **v** moving at right angles to a uniform field **B**, as shown in Fig. 5-39. By equating the force $\mathbf{F} = e\mathbf{v} \times \mathbf{B}$ exerted by the field to the centripetal force, show that the path is circular, with a radius given by

$$r = mv/eB$$

(c) Show that the electron moves around this circle with an angular velocity

$$\omega = eB/m \qquad (5\text{-}52)$$

called the *cyclotron frequency*.

(d) Estimate ω for $m_e = 0.1$ m and $B = 0.5$ weber/m².

For very pure crystals at low temperatures, an electron (or a hole) will be able to travel this circular path without being deflected either by an impurity atom or by the vibrations of the lattice atoms. If an alternating electric field, whose frequency is determined by Problem 5-47(d), is applied, the circulating electron can absorb energy. If we vary the magnetic field, the absorption will show peaks at values of energy corresponding to the various values of the effective masses of both electrons and holes. Figure 5-40, taken from Dresselhaus, Kip, and Kittel (1955),[18] shows two values for m_h and two for m_e. This large number of peaks is seen because the sample had its (100) axis at a 30-deg angle to **B**. By measuring m_e as a function of angle, it was possible to determine that $m_t = 0.08$ m, $m_l = 1.58$ m. It is also possible to determine the dependence of the effective mass on crystallographic direction by rotating the crystal with respect to the magnetic field. This *mass anisotropy* will be considered further in Chapter 7.

The situation at the maximum in the valence band is considerably different. The bands for both germanium and silicon have their maximum at $\mathbf{k} = (000)$ and there is a three-fold degeneracy there. Since the energy is a multivalued function of k, we cannot use a Taylor expansion, since the derivatives do not have a unique definition.

PROBLEM 5-48 Prove that if the energy curves at the top of the valence band for $k = 0$ were non-degenerate, the constant-energy surfaces would be spheres.

For the actual band structure shown in Fig. 5-37, the energy surfaces have an equation.

$$E = -(\hbar^2/2m)[Ak^2 \pm \sqrt{B^2k^4 + C^2(k_x^2k_y^2 + k_y^2k_z^2 + k_z^2k_x^2)}] \qquad (5\text{-}53)$$

where $A = 13, B = 8.7, C = 11.4$ for germanium

$B = 4.1, B = 1.4, C = 3.7$ for silicon

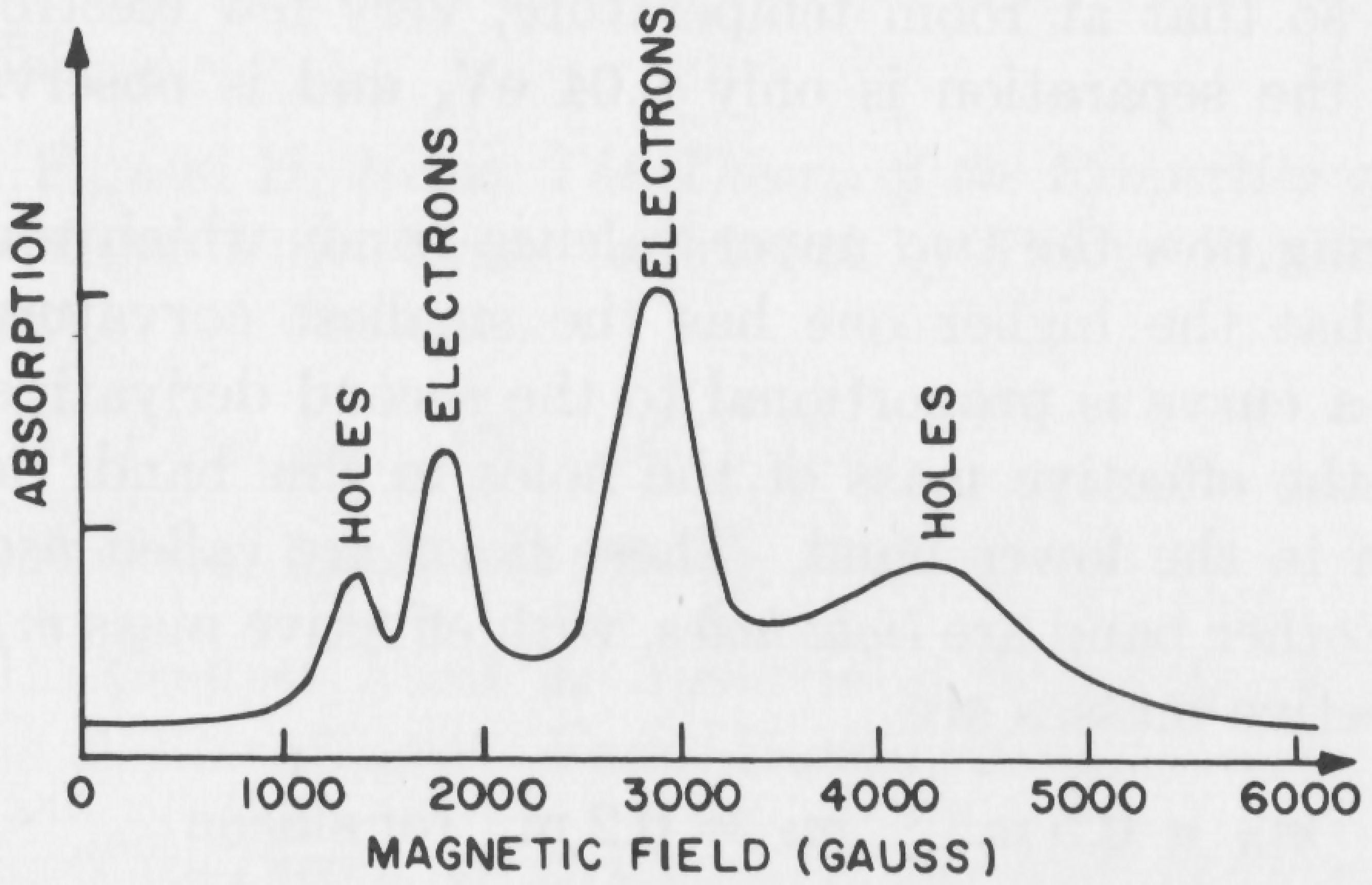


FIG. 5-40. Microwave resonance absorption in germanium. (From G. Dresselhaus et al, Phys. Rev. **98**, 368, 1955, with permission.)

these values being taken from Lax and Mavroides (1955).[19] The derivation of this equation is based on advanced applications of group theory and will not be given. However, in Chapter 7 we shall consider its experimental confirmation. Equation (5-53) can be shown to represent *warped spheres;* that is, spheres which are punched in or pulled out at various points related by the symmetry of the Brillouin zone. Cross sections of the two spheres (one for the positive sign and one for the negative sign) are shown for germanium in Fig. 5-41.

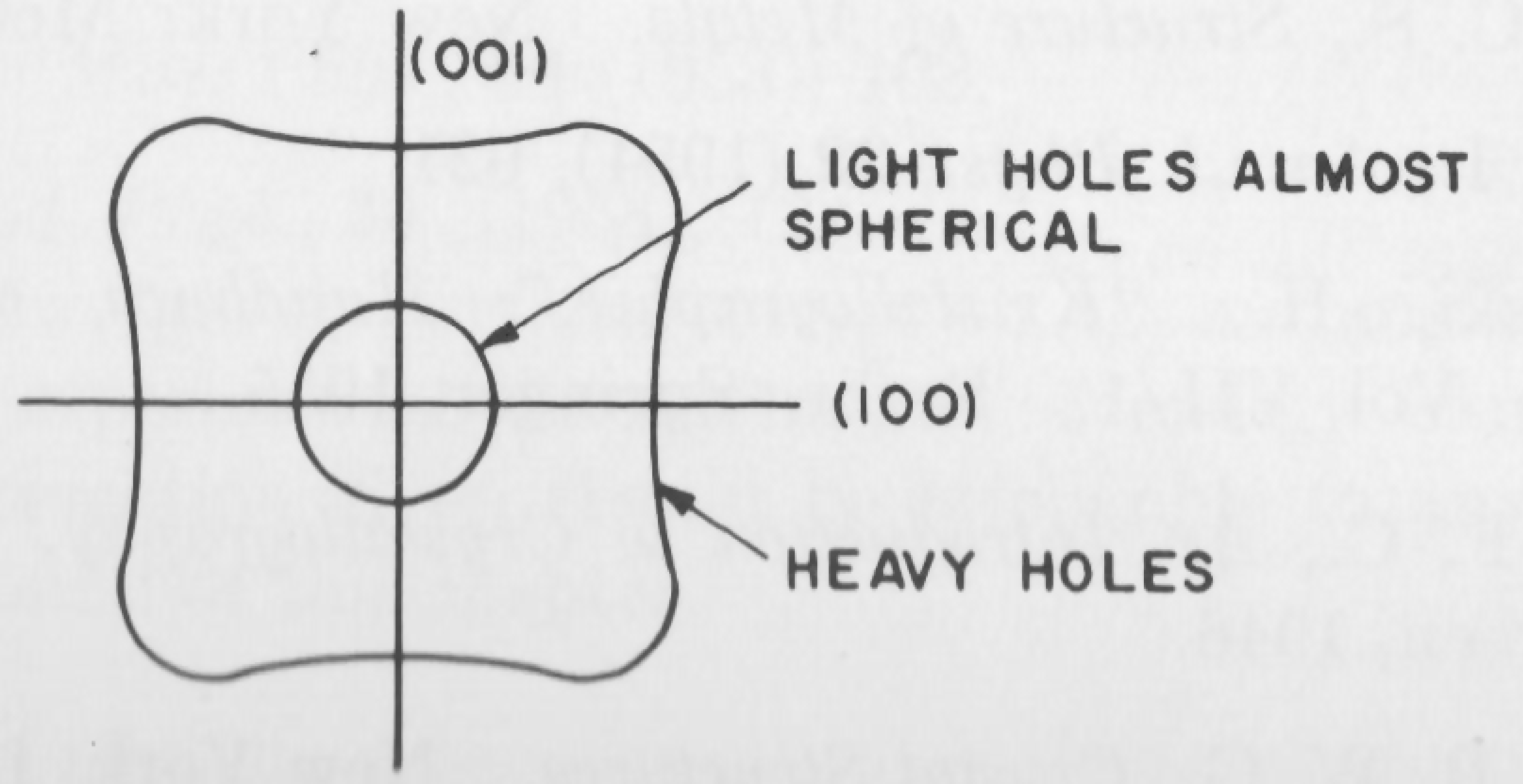


FIG. 5-41. The light and heavy hole bands for silicon. (From *Semiconductors*, ed. N. B. Hannay, Reinhold Publ. Co., New York, 1959, with permission.)

Figures 5-40b and 5-40d represent the changes in the upper figures when spin-orbit splitting is taken into account. When the energy associated with these forces is incorporated into the secular equation, the E vs $\mathbf{k}$ curves are altered from the upper to the lower set of the figure. The effect is to separate the upper, doubly-degenerate valence band for both silicon and germanium into two bands which are degenerate only at $\mathbf{k} = 0$, and to lower the third band. The third band lies 0.28 eV below the other two in germanium, so that at room temperature, very few electrons can leave. For silicon, the separation is only 0.04 eV, and is observable at room temperature.

Considering now the two upper valence bands which touch at $\mathbf{k} = 0$, we notice that the higher one has the smallest curvature. Since this property of a curve is proportional to the second derivative, Eq. (1-116) shows that the effective mass of the holes in this band, denoted m_H, is greater than in the lower band. These holes are called *heavy holes*, and those in the other band are *light holes*, with effective mass m_L. The values of these effective masses are

$$m_H = 0.5\text{ m} \qquad m_L = 0.2\text{ m} \quad \text{for silicon}$$

$$m_H = 0.3\text{ m} \qquad m_L = 0.04\text{ m} \quad \text{for germanium}$$

In Fig. 5-41, the constant-energy surfaces corresponding to light-holes and heavy-holes are indicated.

With the background we have just given on band theory, we can now consider the electrical properties of semiconductors and the interpretation of these properties in terms of this theory.

REFERENCES

1. Bond, W. L., Univ. of Toronto Studies, Geological Series, No. 51 (1947), 9.
2. Barrett, C. S., *Structure of Metals*. New York: McGraw-Hill, 1952.
3. Schiff, L. I., *Am. J. Phys.*, **22** (1954), 621.
4. Jagodzinski, H., *"Kristallographie," Handbuch der Physik*, ed. S. Flügge, Vol. VII/1. Berlin: Springer, 1955.
5. Phillips, F. C., *An Introduction to Crystallography*. London: Longmans, Green, 1946.
6. Wyckoff, R. W. G., *Crystal Structures*. New York: Interscience Publishers, 1960.
7. Davies, T. J., *J. Appl. Phys.*, **28** (1957), 1217.

8. Nomura, K. C., *Phys. Rev. Lett.*, **5** (1960), 500.
9. Beamer, W. and L. R. Maxwell, *J. Chem. Phys.*, **14** (1946), 569.
10. Bradley, A. J., *Phil. Mag.*, **48** (1924), 477.
11. Dekker, A. J., *Solid State Physics*. Englewood Cliffs, N. J.: Prentice-Hall, Inc., 1957.
12. Wilson, A. H., *The Theory of Metals*. Cambridge: Cambridge University Press, 1953.
13. Mott, N. F., and H. Jones, *The Theory of the Properties of Metals and Alloys*. Oxford Univ. Press (reprinted by Dover Publications, New York, 1958).
14. Reitz, J. R., *Phys. Rev.*, **105** (1957), 1233.
15. Herman, F., *Proc. I.R.E.*, **43** (1955), 1703.
16. Mariot, L., *Groupes Finis de Symétrie et Recherche de Solutions de L'Équation de Schrödinger*. Paris: Dunod, 1959.
17. Jones, H., *The Theory of Brillouin Zones and Electronic States in Crystals*. Amsterdam: North-Holland Publishing Co., 1960.
18. Dresselhaus, G., A. F. Kip, and C. Kittel, *Phys. Rev.*, **98** (1955), 368.
19. Lax, B. and J. G. Mavroides, *Phys. Rev.*, **100** (1955), 1650.

GENERAL REFERENCES

Holden, A. and P. Singer, *Crystals and Crystal Growing*. Garden City, New York: Doubleday and Co., 1960. This delightful little book gives a simple discussion of the symmetry properties of crystals and shows how to make paper models of the 32 crystal classes. There are some excellent photographs and instructions on how to grow crystals at home.

Herman, F., *Rev. Mod. Phys.*, **30** (1958), 102.

Lax, B., *Rev. Mod. Phys.*, **30** (1958), 122.

This set of two review articles covers in an authoritative way the theoretical and experimental determination of energy band structures. Most of the information given should be intelligible to anyone who has digested the material of this chapter.

6

IRREVERSIBLE THERMODYNAMICS

6-1 The Onsager Principle

In Chapter 2 we discussed the subject of thermostatics, which deals with matter in thermal, chemical, mechanical, and electrical equilibrium. We now wish to extend some of these ideas to nonequilibrium situations, and, in particular, to consider irreversible processes, which were defined in Chapter 2. Examples of such processes are the flow of electric and thermal currents, governed by the two well-known equations: Ohm's law

$$J = -\sigma \frac{d\mathcal{V}}{dx} \qquad (6\text{-}1)$$

and Fourier's law $J_Q = -\kappa \dfrac{dT}{dx}$ (6-2)

where J is the electrical current density (amperes

per meter^2), J_Q is the thermal-energy current density (watts per meter^2), σ is the electrical conductivity, κ is the thermal conductivity, $d\mathcal{V}/dx$ is the electrical potential gradient, and dT/dx is the thermal gradient. If a body is subject to both electrical and thermal gradients, then Eqs. (6-1) and (6-2) can be generalized to express the fact that when electrons (or holes) are placed in motion by either type of force, a simultaneous transfer of charge and thermal energy occurs. Hence, these equations become

$$J = L_{11} \frac{d\mathcal{V}}{dx} + L_{12} \frac{dT}{dx} \qquad (6\text{-}3)$$

$$J_Q = L_{21} \frac{d\mathcal{V}}{dx} + L_{22} \frac{dT}{dx} \qquad (6\text{-}4)$$

where we have written L_{11} for $-\sigma$ and L_{22} for $-\kappa$, and where L_{12} and L_{21} represent the coupling between the electrical and thermal processes. Equations (6-3) and (6-4) can be written in the more general form

$$J_i = \sum_j L_{ij} F_j \qquad (6\text{-}5)$$

where the F_j represent the driving gradients or *forces*, and the J_i are also known as the *fluxes*. We now wish to prove that when the J's and the F's are properly chosen, then

$$L_{ij} = L_{ji} \qquad (6\text{-}6)$$

This is known as *Onsager's theorem;* it should be emphasized that Eq. (6-6) does *not* hold when the L_{ij} are those of Eqs. (6-3) and (6-4).

To demonstrate the plausibility of Eq. (6-6), we first express the current densities J_i in another form. For irreversible processes, the thermostatic variables U, V, N, etc. will have values different from their equilibrium values U_0, V_0, N_0, etc., and we denote the deviations by α_i, where

$$\alpha_1 = U - U_0, \qquad \alpha_2 = V - V_0, \ldots \qquad (6\text{-}7)$$

The α_i change with time, and the currents J_i can be identified with the derivatives $d\alpha_i/dt$ (abbreviated as $\dot{\alpha}_i$), if the proper dimensionality constants are incorporated into the α_i. For example, from the definition of electric current, we know that J is proportional to $d(q - q_0)/dt$.

We will then replace Eq. (6-5) by a new relation

$$J_i = \dot{\alpha}_i = \sum_j L_{ij} F_j \qquad (6\text{-}8)$$

which appears to be the same, but differs in fact from Eq. (6-5), because the J_i are now defined as time-derivatives of the α_i rather than as the quantities which appear on the right-hand sides of Eqs. (6-1) and (6-2). This means that the L_{ij} and the F_j have a new form, also, which we shall consider shortly. It should be noted that Eq. (6-8) implies that the fluxes are linear functions of the forces, an assumption which is one of the foundations of irreversible thermodynamics.

Let us now recall the statistical definition of entropy, given in Chapter 2

$$S = k \ln W \tag{6-9}$$

where k is Boltzmann's constant and W is the thermodynamic probability. If we now define $P(\alpha_1, \alpha_2, \ldots)\,d\alpha_1\,d\alpha_2 \ldots$ as the probability that the value of α_1 lies between α_1 and $\alpha_1 + d\alpha_2$, the value of α_2 simultaneously lies between α_2 and $\alpha_2 + d\alpha_2$, and so forth, then, in analogy with Eq. (6-9), we can write $P(\alpha_i)$ as

$$P(\alpha_i)\,d\alpha_1\,d\alpha_2 \ldots = e^{S/k}\,d\alpha_1\,d\alpha_2 \ldots \tag{6-10}$$

We now define the *average* of the quantity $\alpha_i F_j$ by

$$< \alpha_i F_j > \; = \int_{-\infty}^{\infty} \alpha_i F_j P(\alpha_i)\,d\alpha_1\,d\alpha_2 \ldots \tag{6-11}$$

PROBLEM 6-1 Demonstrate that the definition given by Eq. (6-11) is a more general form of the arithmetical average.

Differentiating Eq. (6-10) with respect to α_i shows that

$$\frac{\partial S}{\partial \alpha_i} = \frac{k\partial P/\partial \alpha_i}{P} \tag{6-12}$$

In Eq. (6-8) we defined the J_i and the L_{ij} in terms of the thermodynamic variables. We now similarly define the F_i by the relation

$$F_i = \partial S/\partial \alpha_i \tag{6-13}$$

Then Eqs. (6-12) and (6-13), substituted into Eq. (6-11), give

$$< \alpha_i F_j > \; = k \int \alpha_i (\partial P/\partial \alpha_j)\,d\alpha_1\,d\alpha_2 \ldots \tag{6-14}$$

When we integrate by parts, the integrated term vanishes at the limits, since the probability of an infinite deviation is zero. Hence,

$$\begin{aligned} < \alpha_i F_j > \; &= -k \int P\,\delta_{ij}\,d\alpha_1\,d\alpha_2 \ldots \\ &= -k\,\delta_{ij} \end{aligned} \tag{6-15}$$

where δ_{ij} is the Kronecker delta defined in Appendix A-2.

Next, let us consider the product of the value of one of the α's, say α_i, at a time t, with another, α_j, at a time τ later. We will denote the average of all such products by $< \alpha_i(t)\alpha_j(t + \tau) >$. If we assume now that the total energy of system in which the irreversible process is occurring depends only on the positions and the squares of the velocities of the charged particles involved, then this average $< \alpha_i(t)\alpha_j(t + \tau) >$ should be independent of a reversal of all velocities. Another way of saying this is that we can reverse the time without affecting the average, or

$$< \alpha_i(t)\alpha_j(t + \tau) > \; = \; < \alpha_i(t)\alpha_j(t - \tau) > \tag{6-16}$$

One exception to this occurs when magnetic fields are present.

PROBLEM 6-2

(a) To what power of the velocity is the force on a charged particle moving in a magnetic field proportional?

(b) Apply the results of part (a) to prove the statement made immediately above.

If τ is a small interval, we can replace t by $(t + \tau)$ in the left-hand side of Eq. (6-16). Then we subtract $< \alpha_i(t)\alpha_j(t) >$ from both sides and divide by τ to get

$$< \dot{\alpha}_i(t)\alpha_j(t) > \; = \; < \alpha_i(t)\dot{\alpha}_j(t) > \tag{6-17}$$

Substituting Eqs. (6-8) and (6-15) into this equation proves Eq. (6-6).

PROBLEM 6-3 Verify the steps that lead from Eq. (6-17) to Eq. (6-6).

6-2 Conjugate Forces and Fluxes

In order to apply the Onsager theorem, we must know how to make the correct selection of the forces F_i and fluxes J_i, in Eq. (6-8), and we shall now show how this is done. Let us start with the relation

$$dU = T\,dS + \bar{\mu}\,dN \tag{6-18}$$

which may be obtained from the *VATVUS* diagram, Fig. 2-2. This equation is valid only for equilibrium conditions, but it can be applied to the non-equilibrium case by associating with every infinitesimal region a local entropy $S(U, V, N_1, \ldots)$ and defining the functional dependence of S on U, V, $N_1, \ldots$ as being identical

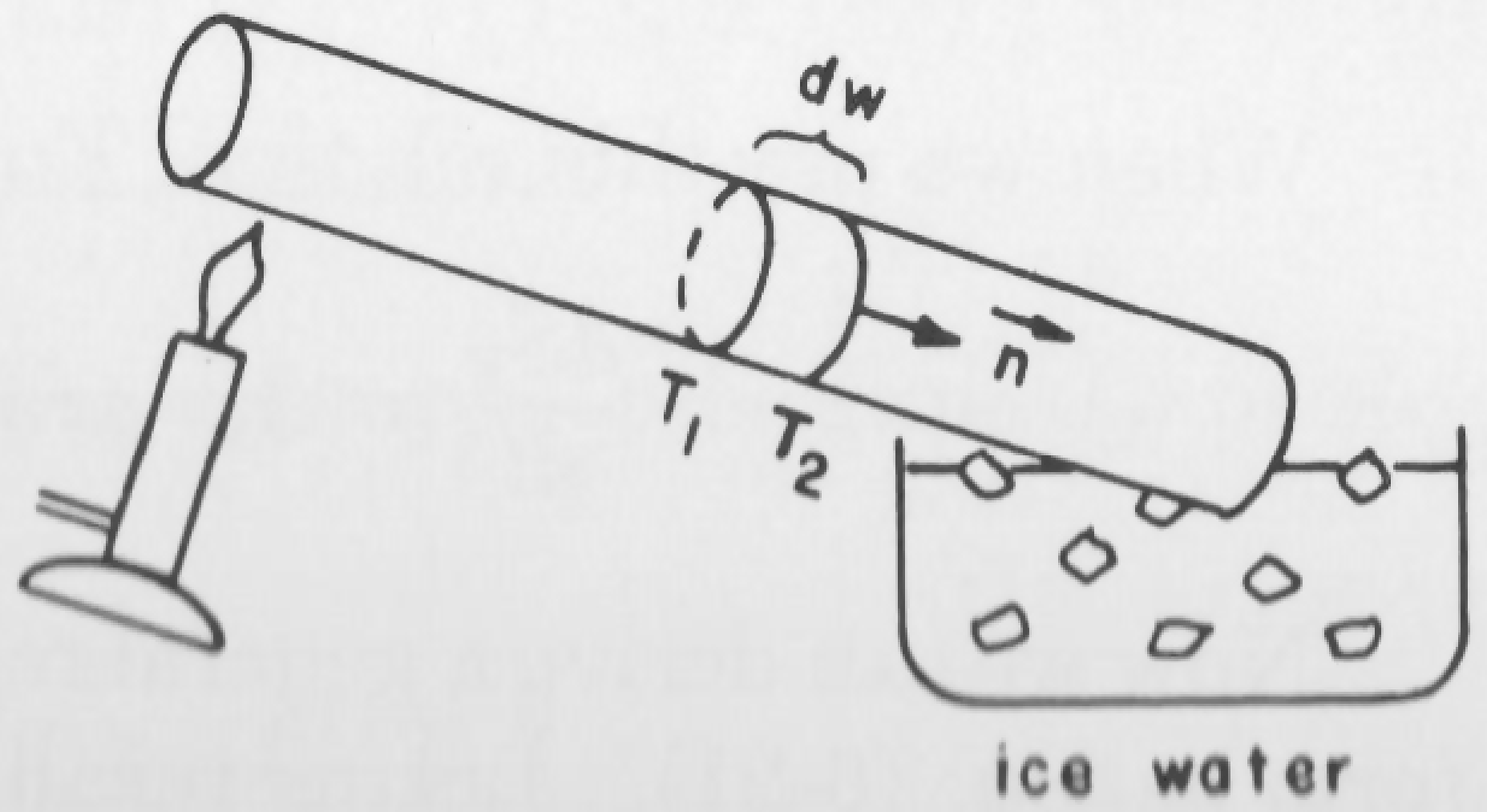


FIG. 6-1. The transport of heat energy.

to that for equilibrium. Figure 6-1 shows a bar carrying heat energy; we have marked an infinitesimal region of width dw whose hot end is at temperature T_1 and whose cold end is at T_2. The currents or fluxes through this volume element can be defined in terms of the quantities in Eq. (6-18), as follows:

$$\mathbf{J}_S = (dS/dt\,dA)\mathbf{n} = \textit{entropy current density}$$

$$\mathbf{J}_U = (dU/dt\,dA)\mathbf{n} = \textit{energy current density}$$

$$\mathbf{J}_N = (dN/dt\,dA)\mathbf{n} = \textit{particle current density}$$

where dA is an element of area and $\mathbf{n}$ is a unit vector normal to dA. Equation (6-18) then becomes

$$\mathbf{J}_U = T\,\mathbf{J}_S + \bar{\mu}\,\mathbf{J}_N \qquad (6\text{-}19)$$

We can take the divergence of both sides of this equation by using the vector identity

$$\operatorname{div} A\mathbf{B} = A \operatorname{div} \mathbf{B} + \mathbf{B}\cdot \mathbf{grad}\, A \qquad (\text{A-33})$$

We realize that $\operatorname{div} \mathbf{J}_U = 0$ when we consider the physical meaning of the divergence, as discussed in the appendix, for the net flow of energy over the surface of the cylinder of width dw must vanish. This is due to the conservation of energy, and, similarly, $\operatorname{div} \mathbf{J}_N = 0$ due to particle conservation. However, the situation is different for $\mathbf{J}_S$. At temperature T_1, an amount of heat dQ brings an amount of entropy $dS_1 = dQ/T_1$ into the infinitesimal volume, and an amount $dS_2 = dQ/T_2$ leaves. Since $T_1 > T_2$, then $dS_2 > dS_1$, and there is, consequently, an irreversible generation of entropy within each elementary volume of the bar. The *generation-rate of entropy per unit volume* is denoted by dS_V/dt; again, using the physical meaning of the divergence, we know that

$$\frac{dS_V}{dt} = \operatorname{div} \mathbf{J}_S \qquad (6\text{-}20)$$

PROBLEM 6-4 By considering an amount of entropy ΔS generated in a time Δt in the volume element shown in Fig. 6-1, establish Eq. (6-20).

When we use this relation, Eq. (6-20) becomes

$$\frac{dS_V}{dt} = \mathbf{J}_U\cdot\mathbf{grad}\,\frac{1}{T} - \mathbf{J}_N\cdot\mathbf{grad}\left(\frac{\bar{\mu}}{T}\right) \qquad (6\text{-}21)$$

Now we can derive a general relation for dS_V/dt which will be of the same form as Eq. (6-21). Let us recall that we defined the deviations in U, V, $N, \ldots$ by $\alpha_1 = U - U_0$, $\alpha_2 = V - V_0, \ldots$. When any of these variables U, V, $\ldots$ deviate from their equilibrium value, then S will, similarly, deviate from S_0. The dependence of S on the other variables can then be expressed

$$S = S(\alpha_1, \alpha_2, \ldots)$$

Hence,

$$dS = \frac{\partial S}{\partial \alpha_1}\,d\alpha_1 + \frac{\partial S}{\partial \alpha_2}\,d\alpha_2 + \cdots$$

$$= \sum_i \frac{\partial S}{\partial \alpha_i}\,d\alpha_i$$

and

$$\frac{dS}{dt} = \sum_i \frac{\partial S}{\partial \alpha_i}\,\dot{\alpha}_i \qquad (6\text{-}22)$$

By Eqs. (6-8) and (6-13), this becomes

$$\frac{dS_V}{dt} = \sum_i F_i J_i \qquad (6\text{-}23)$$

which is the desired relation.

PROBLEM 6-5 Explain how dS/dt becomes dS_V/dt in going from Eq. (6-22) to Eq. (6-23).

Comparing Eqs. (6-21) and (6-23), we see that $\mathbf{J}$, $\mathbf{grad}\,(1/T)$ and $-\mathbf{J}_N$, $\mathbf{grad}\,(\bar{\mu}/T)$ form *conjugate pairs* of fluxes and forces. The equations of the form of Eqs. (6-3) and (6-4) are then

$$\mathbf{J}_U = L_{UU}\,\mathbf{grad}\,\frac{1}{T} + L_{UN}\,\mathbf{grad}\,\frac{\bar{\mu}}{T} \qquad (6\text{-}24)$$

$$\mathbf{J}_N = L_{NU}\,\mathbf{grad}\,\frac{1}{T} + L_{NN}\,\mathbf{grad}\,\frac{\bar{\mu}}{T} \qquad (6\text{-}25)$$

with

$$L_{UN} = L_{NU} \qquad (6\text{-}26)$$

PROBLEM 6-6 Are the above equations valid if $(1/T)$ and $(\bar{\mu}/T)$ are interchanged?

We can find other conjugate pairs. For example, substitute J_U of Eq. (6-19) into Eq. (6-21), getting

$$\frac{dS_V}{dt} = -\mathbf{J}_S\cdot\frac{\mathbf{grad}\,T}{T} - \mathbf{J}_N\cdot\frac{\mathbf{grad}\,\bar{\mu}}{T}$$

If we define the heat current density $\mathbf{J}_Q$ as $T\mathbf{J}_S$ in analogy with $dQ = TdS$ and the electric current density as $\mathbf{J} = e\mathbf{J}_N$, this becomes

$$\frac{dS_V}{dt} = -\mathbf{J}_Q\cdot\mathbf{grad}\,\frac{1}{T} - \frac{\mathbf{J}}{e}\cdot\frac{\mathbf{grad}\,\bar{\mu}}{T} \tag{6-27}$$

so the generalized Ohm and Fourier laws in Onsager form are

$$-\mathbf{J}/e = L_{11}\frac{\mathbf{grad}\,\bar{\mu}}{T} + L_{12}\,\mathbf{grad}\,\frac{1}{T} \tag{6-28}$$

$$\mathbf{J}_Q = L_{21}\frac{\mathbf{grad}\,\bar{\mu}}{T} + L_{22}\,\mathbf{grad}\,\frac{1}{T} \tag{6-28}$$

with

$$L_{12} = L_{21} \tag{6-30}$$

We can now relate the coefficients L_{ij} to familiar constants. First, we recall that the electrical conductivity σ is usually defined by the equation

$$\sigma = J/\mathcal{E} = -J/\text{grad}\,\mathcal{V} \tag{6-31}$$

where $\mathcal{E}$ is the electric field strength and $\mathcal{V}$ is the electric potential. For the situation where particle and temperature gradients can exist, we generalize this by defining σ as the ratio.

$$\sigma = \left.\frac{-J}{(1/e)\,\text{grad}\,\bar{\mu}}\right)_{\text{grad}\,T=0} \tag{6-32}$$

By Eq. (6-28), this becomes

$$\sigma = e^2L_{11}/T \tag{6-33}$$

Similarly, we define the thermal conductivity κ as

$$\kappa = -\left.\frac{J_Q}{\text{grad}\,T}\right)_{J=0} \tag{6-34}$$

From this it follows that

$$\kappa = D/L_{11}T^2 \tag{6-35}$$

where D is the determinant $L_{11}L_{22} - L^2_{12}$.

6-3 Thermoelectric Power

Consider bare wires of two different kinds, A and B, joined at two points, as shown in Fig. 6-2. The junctions are maintained at temperatures T_1 and T_2, and wire B is cut and has a voltmeter connected as shown. It is found that a potential difference will be developed across the voltmeter, whose value depends on the temperature difference $(T_1 - T_2)$. This phenomenon is called the *Seebeck effect*, and we define the *Seebeck coefficient* or *relative thermoelectric power* ϵ_{AB} of the two metals as the rate of change of potential difference with temperature difference. For simplicity, we will refer to ϵ_{AB} as the *relative thermopower*. We can also speak of the *absolute thermopower* ϵ of a single substance, for if we take a piece of wire and heat it locally, electrons near the source of heat will acquire kinetic energy and move away. This sets up an electrical potential gradient due to the temperature gradient. Hence, we define the absolute thermopower as

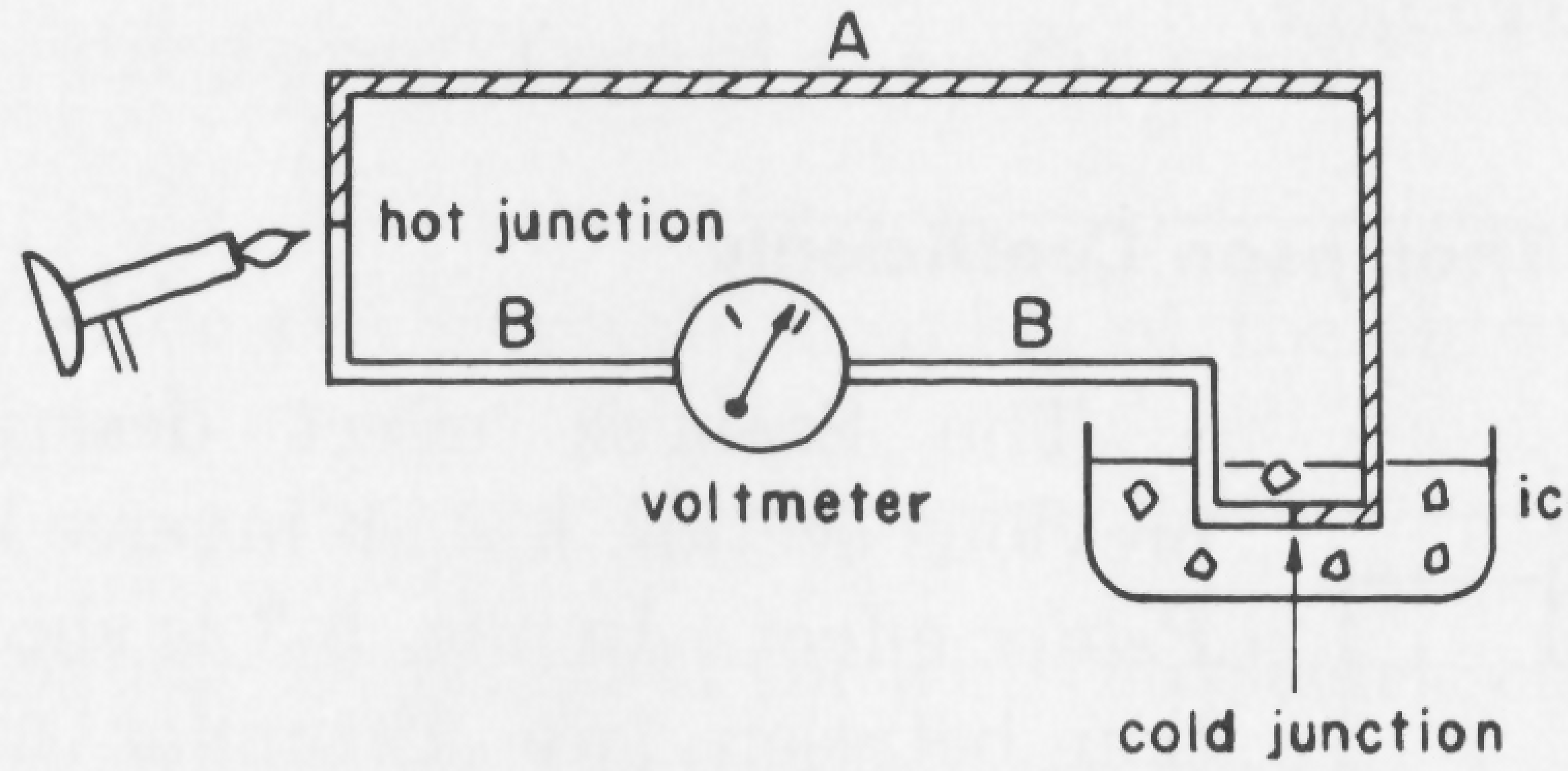


FIG. 6-2. A thermocouple circuit.

$$\epsilon = -\left.\frac{\mathbf{grad}\,\bar{\mu}}{e\,\mathbf{grad}\,T}\right)_{J=0} \tag{6-36}$$

From Eq. (6-28)

$$\epsilon = -L_{12}/eTL_{11} \tag{6-37}$$

Using this and Eqs. (6-33) and (6-28) gives the *generalized Ohm's Law*

$$-\mathbf{J} = \sigma\epsilon\,\mathbf{grad}\,T + (\sigma/e)\,\mathbf{grad}\,\bar{\mu} \tag{6-38}$$

PROBLEM 6-7 What circumstances will cause Eq. (6-38) to reduce to the simple form of Ohm's law

$$\mathbf{J} = \sigma\mathcal{E} \tag{6-39}$$

In the same way the L_{ij} can be eliminated from Eq. (6-29) so that both Eq. (6-28) and Eq. (6-29) can be written in terms of ϵ, κ, and σ.

PROBLEM 6-8

(a) Eliminate L_{21} and L_{22} from Eq. (6-29) to get

$$\mathbf{J}_Q = -(\epsilon T/e)\,\mathbf{grad}\,\bar{\mu} - (T\sigma\epsilon^2 + \kappa)\,\mathbf{grad}\,T \tag{6-40}$$

(b) Show that

$$\mathbf{J}_S = e\epsilon\mathbf{J}_N - (\kappa/T)\,\mathbf{grad}\,T \tag{6-41}$$

According to Eq. (6-41), the total entropy current consists of two terms.

The second term in Eq. (6-41), by Eq. (6-2), is the heat current $\mathbf{J}_Q$ divided by T; the first term is the entropy carried by the motion of charges and indicates that an amount of entropy $e\epsilon$ is carried by each charged particle. Hence, we can regard the thermopower as the entropy transported per coulomb by the charge flow.

6-4 The Peltier and Thompson Coefficients

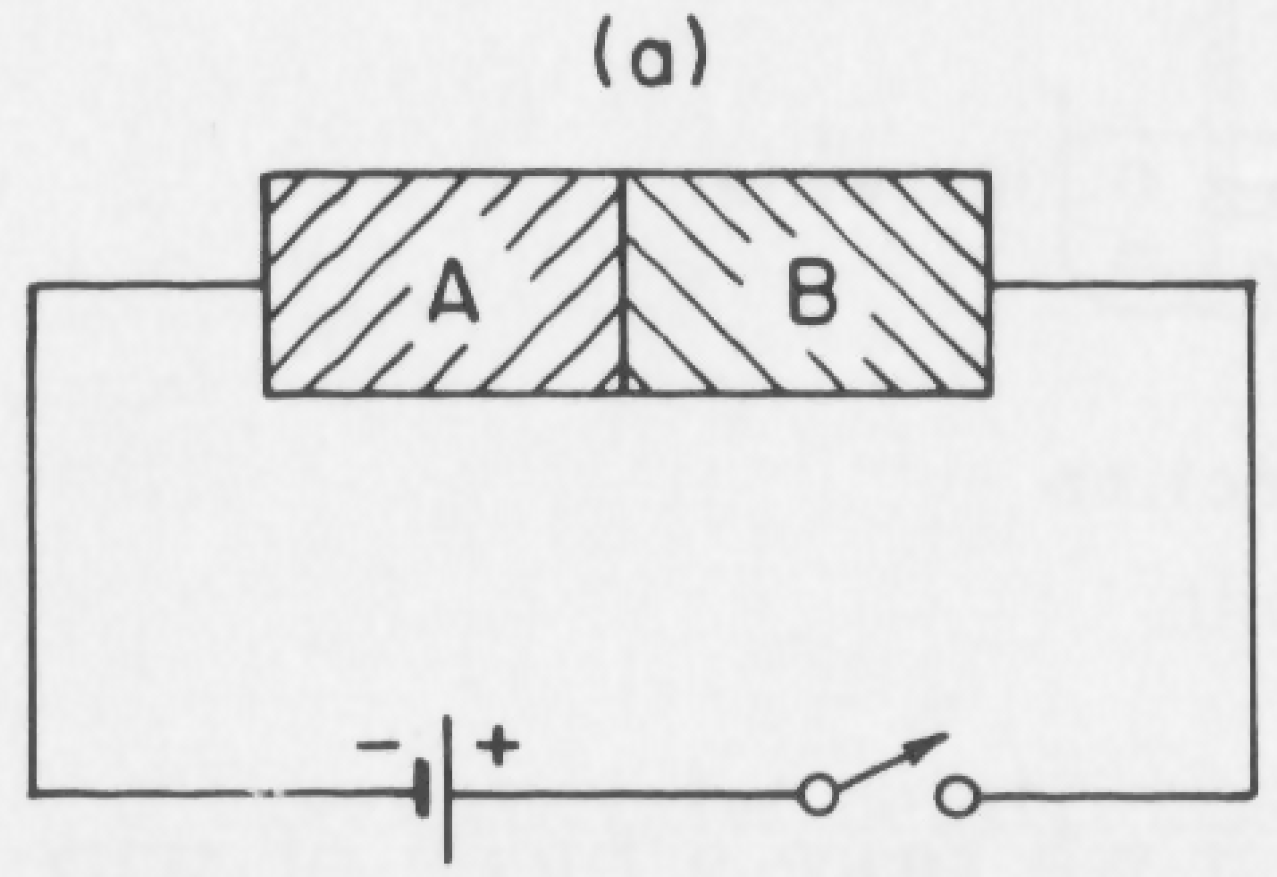


FIG. 6-3. A circuit for demonstrating the Peltier effect; A and B are dissimilar metals.

The Seebeck effect described in the previous section has an inverse known as the *Peltier* effect. In Fig. 6-3 is shown the junction between two dissimilar metals. It is found that when the switch is closed and current flows through the junction, either heating or cooling will occur, and that the heating will change to cooling, or vice versa, when the current is reversed. Another way of describing this effect is to say that when an electric current $\mathbf{J}$ is flowing, a heat current $\mathbf{J}_Q$ must enter or leave the system in order to maintain the temperature constant. It is found experimentally that $\mathbf{J}_Q$ is proportional to $\mathbf{J}$, or

$$J = \Pi_{AB} J_Q \tag{6-42}$$

where Π_{AB} is the *relative Peltier coefficient.*

By Eqs. (6-28) and (6-29), when $\mathbf{grad}\ T = 0$,

$$J_Q/J = L_{21}/eL_{11} \tag{6-43}$$

so that Eqs. (6-42) and (6-43) give

$$\Pi_{AB} = -L_{21}/eL_{11} \tag{6-44}$$

Using the Onsager theorem and Eq. (6-37), we obtain

$$\Pi_{AB} = \epsilon_{AB} T \tag{6-45}$$

This equation, connecting the Peltier and Seebeck coefficients, is known as *Kelvin's second relation,* and has been verified experimentally. Hence, it can be considered as a confirmation of the Onsager theorem.

Next, consider a single homogeneous wire of constant cross section in which a temperature gradient exists. If we now permit an electric current to pass through the wire, the temperature distribution would be expected to change, due to the evolution or absorption of heat. This is called the *Thompson effect;* we will now derive an expression for the Thompson heat. By Eq. (6-19),

$$\text{div}\ \mathbf{J}_U = \text{div}\ \mathbf{J}_Q + (\bar{\mu}/e)\ \text{div}\ \mathbf{J} \tag{6-46}$$

where $\mathbf{J}_Q = T\mathbf{J}_S$ and $\mathbf{J} = q\,\mathbf{J}_N$. If we use Eqs. (6-38) and (6-41), Eq. (6-46) becomes

$$\text{div}\ \mathbf{J}_U = \text{div}\ (T\epsilon\mathbf{J} - \mathbf{grad}\ T) - \left(\frac{e^2}{\sigma}\mathbf{J}_N + \epsilon\ \mathbf{grad}\ T\right)\cdot\mathbf{J}$$

or

$$\text{div}\ \mathbf{J}_U = T\mathbf{J}\cdot\mathbf{grad}\ \epsilon - \kappa\ \text{div}\ \mathbf{grad}\ T - (J^2/\sigma) \tag{6-47}$$

PROBLEM 6-9 Give the physical meaning of the three terms on the right of Eq. (6-47).

For the original conditions of the wire (no electric current and a constant temperature distribution), the total energy current $\mathbf{J}_U$ entering and leaving any small volume must be the same, or div $\mathbf{J}_U = 0$ when $\mathbf{J} = 0$. This can only be true if div $\mathbf{grad}\ T = 0$ in Eq. (6-47). Hence

$$\text{div}\ \mathbf{J}_U = T\mathbf{J}\cdot\mathbf{grad}\ \epsilon - (J^2/\sigma) \tag{6-48}$$

If we use

$$\mathbf{grad}\ \epsilon = \frac{d\epsilon}{dT}\ \mathbf{grad}\ T \tag{6-49}$$

Eq. (6-48) becomes

$$\text{div}\ \mathbf{J}_U = T\frac{d\epsilon}{dT}\mathbf{J}\cdot\mathbf{grad}\ T - (J^2/\sigma) \tag{6-50}$$

The second term on the right will be recognized as the Joule heat. The first term is the Thompson heat and, in terms of it, we define the *Thompson coefficient* τ as the Thompson heat per unit electric current and per unit temperature gradient

$$\tau = \frac{T\dfrac{d\epsilon}{dT}\mathbf{J}\cdot\mathbf{grad}\ T}{\mathbf{J}\cdot\mathbf{grad}\ T} = T\frac{d\epsilon}{dT} \tag{6-51}$$

Note that the Joule heat depends on J to the second power, and hence is independent of a reversal of the direction of current flow, whereas the Thompson heat will reverse. Equation (6-51) is known as the *first Kelvin relation.*

GENERAL REFERENCES

Kittel, C., *Elementary Statistical Physics.* New York: John Wiley & Sons, 1958. This book gives a clear, concise treatment of irreversible thermodynamics and discusses important applications to semiconductor theory.

Callen, H. B., *Thermodynamics.* New York: John Wiley and Sons, 1960. This book develops an axiomatic approach to thermodynamics which is quite different from previous treatments of the subject. Both thermostatics and irreversible thermodynamics are covered.

Domenicali, C. A., *Rev. Mod. Phys.*, **26** (1954), 237. This review article gives a very lucid and thorough discussion of the experimental aspects of irreversible thermodynamics, particularly with regard to thermoelectricity.

7

TRANSPORT THEORY

7-1 An Introduction to the Galvanomagnetic Coefficients

In the previous chapter, we considered the thermoelectric effects from what is known as a *phenomenological* point of view; that is, we discussed relations like Fourier's and Ohm's laws, which connect the experimentally measured forces and fluxes, but we ignored the mechanisms through which these quantities are related. The object of *transport theory* is to study such mechanisms and to base the calculations of semiconductor properties such as resistivity and thermopower on these processes. The fundamental quantitative relation in this theory is the *Boltzmann transport equation*, whose plausibility is discussed in the next section. There is a tremendous amount of detail involved in setting up and solving this equation and in applying the solutions; it is worthwhile, therefore, to obtain an intuitive insight into the maze of mathematics which follows.

Consider first the electrical properties of a semiconductor. In Chapter 1 we defined the electrical conductivity σ and its reciprocal, the resistivity ρ. When a semiconductor is subject to an electric field $\mathbf{\mathcal{E}}$, the electrons and holes contribute to the current density $\mathbf{J}$, and, in general, $\mathbf{J}$ and $\mathbf{\mathcal{E}}$ are related through Ohm's law.

$$\mathbf{J} = \sigma\mathbf{\mathcal{E}}$$

In Chapter 1 we also defined magnetic induction $\mathbf{B}$ by the relation

$$\mathbf{F} = e\mathbf{v} \times \mathbf{B} \tag{7-1}$$

where $\mathbf{F}$ is the force in newtons on a particle of charge e coulombs moving with velocity $\mathbf{v}$ meters/second in a magnetic field of $\mathbf{B}$ webers/meter^2. Suppose we now apply both an electric field and a magnetic field at right angles to a semiconducting slab, as shown in Fig. 7-1. If we assume for the moment that only electrons are carrying current in the semiconductor, then according to the above equation, electrons moving upward with a velocity $\mathbf{v}$ will experience a force which urges them towards the front edge (since $q = -1.6 \times 10^{-19}$ coulomb), producing a negative charge in this region. By conservation of charge there will be a deficiency of electrons, or a positive charge, on the rear edge, and when we connect a high-impedance voltmeter as shown, we will measure a voltage of the polarity indicated. This voltage is called the *Hall voltage*, and the phenomenon is the *Hall effect* (first reported by E. H. Hall in 1879). The corresponding Hall field $\mathcal{E}_H$ is in a direction which opposes the magnetic force qBv, and when the two forces balance, no further electrons are deflected. If we now replace the N-type semiconductor with a P-type slab, we reverse three effects: the velocity-vector, the direction of deflection, and the polarity of accumulated charge. This reverses the sign of the Hall voltage as read on the meter and provides a very convenient way of determining whether the material is N-type or P-type.

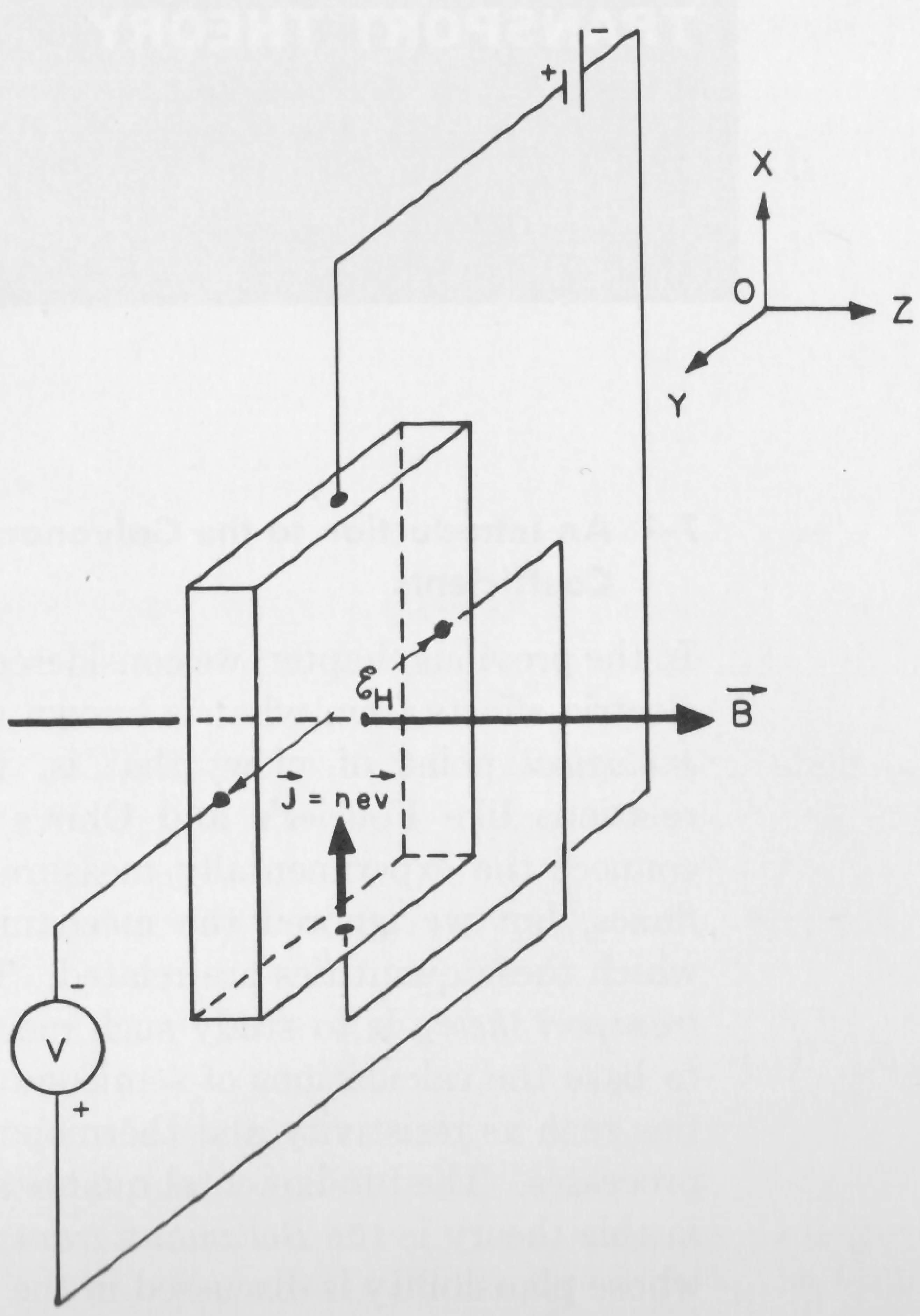


FIG. 7-1. The relation of the electric and magnetic quantities involved in the Hall effect.

Another fundamental electrical property of a semiconductor is the mobility μ of each kind of carrier. We shall consider electrons only, for the moment; we showed in Eqs. (1-22) through (1-24), that

$$J = nev \tag{7-2}$$

and

$$\sigma = ne\mu = nev/\mathcal{E} \tag{7-3}$$

so that

$$\mathbf{J} = \sigma\mathbf{\mathcal{E}} = ne\mu\mathbf{\mathcal{E}} = ne\mathbf{v} \tag{7-4}$$

Electrons moving in a semiconductor collide with the atoms in the crystal lattice, and the motion can no longer be represented by the straight lines we have been using, but instead will be some sort of a zigzag path proceeding in the general direction of the positive electrical lead. For this reason, we refer to the electron as *drifting* in the electric field and call μ the *drift mobility*. If we define the *collision time* τ as the average time between these collisions, then the acceleration a of the electron is approximately given by

$$a = F/m = e\mathcal{E}/m = v/\tau \tag{7-5}$$

By Eq. (1-23),

$$\mu = e\tau/m \tag{7-6}$$

and by Eq. (1-24)

$$\sigma = ne\mu = ne^2\tau/m \tag{7-7}$$

Now suppose τ is not the same for all electrons; let there be n_1 electrons with collision time τ_1, n_2 electrons with collision time τ_2, and so on. Then the conductivity is

$$\sigma = \sum_i \sigma_i = (e^2/m)\sum_i n_i\tau_i \tag{7-8}$$

Multiplying top and bottom by n gives

$$\sigma = (ne^2/m) \sum n_i \tau_i / n \tag{7-9}$$

The quantity $\Sigma\, n_i\tau_i/n$ is the definition of the arithmetical average and is called the *average collision time* $< \tau >$. Hence,

$$\sigma = ne^2 < \tau > / m \tag{7-10}$$

and the corresponding mobility is

$$\mu = e < \tau > / m \tag{7-11}$$

which should be compared with Eq. (7-6).

We can extend this type of calculation to the Hall coefficient. We would expect the Hall field $\mathcal{E}_H$ in Fig. 7-1 to be proportional to both B and J, or

$$\mathcal{E}_H = RBJ \tag{7-12}$$

where R is called the *Hall coefficient*. When the electric and magnetic forces balance,

$$B_z e v_x = e\mathcal{E}_y \tag{7-13}$$

where the subscripts refer to the axes shown in Fig. 7-1. Using Eqs. (7-4) and (7-13) in Eq. (7-12), we obtain

$$R = 1/ne \tag{7-14}$$

which indicates that the MKS units of the Hall coefficient are meter^3/coulomb. (However, cm^3/coulomb are commonly used.) Combining Eqs. (7-12) and (7-14), we see that

$$n = BJ/\mathcal{E}_H e \tag{7-15}$$

which tells us that by measuring the voltage produced by a given current and magnetic field, we can determine the electron concentration n. If we assume that each donor atom contributes one electron, that the number of donors is much larger than the number of acceptors, and that the electrons come from the donors and not from the valence band (which we insure in many materials by cooling the sample to liquid nitrogen), the Hall effect, by Eq. (7-15), provides an electrical method for the determination of impurity concentration. Taking into account refinements to be discussed in a later chapter, this method is known to be accurate to a few parts per billion, far beyond the reach of spectroscopic analysis.

Consider now the case where the electrons do not all have the same velocity and collision time. For any group n_i, by Eqs. (7-5) and (7-13)

$$Bev_{xi} = mv_{yi}/\tau_i$$

or

$$v_{yi} = (eB/m)\tau_i v_{xi} \tag{7-16}$$

Eliminate v_{yi} and v_{xi} by using Eqs. (7-4) and (7-7) to get

$$\sigma_i \mathcal{E}_y = (eB/m)\tau_i \sigma_i \mathcal{E}_x$$

By Eq. (7-8), the right-hand side of this is changed to give

$$\sigma_i \mathcal{E}_y = (eB/m)\tau_i (n_i e^2 \tau_i / m)\mathcal{E}_x$$

Summing both sides gives

$$\sigma \mathcal{E}_y = (e^3 B/m)\mathcal{E}_x n < \tau^2 > = (ne^3 B/m)(J_x/\sigma) < \tau^2 > \tag{7-17}$$

where

$$< \tau^2 > = \sum n_i \tau_i^2 / n$$

Putting Eq. (7-17) into Eq. (7-12) yields

$$R = E_y/J_x B_z = \frac{ne^3 < \tau^2 >}{m^2 \sigma^2} = \frac{1}{ne} \frac{< \tau^2 >}{< \tau >^2} \tag{7-18}$$

This indicates that the simple formula of Eq. (7-14) should be corrected by the term $< \tau^2 > / < \tau >^2$, which incorporates the subtle details of the way the electrons zigzag among the atoms of the lattice. The calculation of $< \tau >$ and $< \tau^2 >$ will be performed later in this chapter, and we shall show as an example that $< \tau^2 > / < \tau >^2 = 3\pi/8 = 1.18$ for the situation in which the energy surfaces are spherical and the scattering is due primarily to the vibrating atoms of the lattice, whereas for scattering by ionized impurities, $< \tau^2 > / < \tau >^2 = 315\pi/512 = 1.93$.

When both electrons and holes are present in a semiconductor, we have a more complicated problem. For electrons, from Eq. (7-13)

$$\mathcal{E}_y = -Bv_{ex}$$

and

$$J_n = ne\,\mu_n\, \mathcal{E}_{\underset{y}{x}} = -ne\,\mu_n^2 B\, \mathcal{E}_x$$

Adding a similar expression for holes, we obtain

$$J_y = Be\mathcal{E}_x(-\,n\mu_n^2 + p\mu_p^2)$$

Using Eq. (7-12) and assuming lattice scattering, we obtain

$$R = \frac{3\pi(p\mu_p^2 - n\mu_n^2)}{8e(ne\mu_n + pe\mu_p)} \tag{7-19}$$

If we let $b = \mu_n/\mu_p$ denote the *mobility ratio*, this becomes

$$R = 3\pi(p - nb^2)/8e(p + nb)^2 \tag{7-20}$$

Note that when either n or p vanishes, Eq. (7-20) reduces to the previous expression (with the proper sign).

We can interpret the Hall coefficient geometrically in terms of an angle, the *Hall angle* θ, which can be identified with the deflection angle in Fig. 5-39. The Hall angle is determined by the circular path the electron traverses before it is scattered by a collision with an atom. The time corresponding to this path is the collision time τ, and θ and τ are related by

$$\theta = \omega\tau = eB\tau/m \tag{7-21}$$

The size of the Hall angle would be expected to be a property of the material, and we define the *Hall mobility* μ_H as the Hall angle per unit magnetic field

$$\mu_H = \theta/B = e\tau/m \tag{7-22}$$

Note that the definitions of μ and μ_H are analogous and that the formulas of Eqs. (7-22) and (7-6) are similar, although the τ's are not necessarily the same. From Eqs. (7-7) and (7-14), we have another important expression for the Hall mobility.

$$\mu_H = R\sigma \tag{7-23}$$

If we use Eqs. (7-11) and (7-18) in connection with this definition, then

$$\mu_H = (e/m)(<\tau^2>/<\tau>) \tag{7-24}$$

which is dimensionally the same as Eq. (7-11), but which represents a different kind of average

PROBLEM 7-1 Prove that

$$\theta = \mathcal{E}_y/\mathcal{E}_x \tag{7-25}$$

The presence of the magnetic field is responsible for another important effect, called the *magnetoresistance*. Glicksman (1958)[1] has given a simple interpretation of this phenomenon. Consider first a group of electrons all moving with the same velocity. When a magnetic field is applied, the deflecting force veB will be balanced by the transverse force $e\mathcal{E}_H$ of Fig. 7-1, and once this equilibrium is established, no further deflection occurs. If there is a distribution of velocities, those electrons whose velocity is precisely that determined by Eq. (7-13) will be undeflected, whereas those whose velocity is above or below will be deflected in opposite directions with respect to the long axis of the crystal. These two streams of electrons produce transverse currents with which are associated magnetic fields. The resulting forces retard the motion of the conduction electrons, producing a rise in resistivity. We define the *magnetoresistance coefficient* M as the ratio of the change in resistivity to the resistivity with no magnetic field

$$M = \frac{\rho - \rho_0}{\rho_0} = \frac{\Delta\rho}{\rho_0} \tag{7-26}$$

Since the size of Hall effect is directly proportional to the value of B, we would expect the magnetoresistance, which depends on the secondary magnetic field, to vary as B^2.

The magnetoresistivity can be treated in the same intuitive fashion as the resistivity, Hall coefficient, and the two mobilities. From Eq. (7-26), we can define the *transverse magnetoresistivity* M_t by

$$M_t = (\rho - \rho_0)/\rho_0 = \Delta\rho/\rho_0$$

where ρ is the resistivity with the magnetic field **B** applied in the transverse direction to **J**, as in Fig. 7-1, and ρ_0 is the resistivity with no field. When the field is off, the current **J** flows parallel to the x-axis, but when **B** is on, **J** has components along OX and OY. By Eq. (7-25)

$$\mathcal{E}_y = \omega\tau_i\,\mathcal{E}_x$$

or

$$J_{yi} = \omega\tau_i J_{xi}$$

Then

$$J_i = (J_{xi}^2 + J_{yi}^2)^{1/2} = J_{xi}(1 + \omega^2\tau_i^2/2)$$

where we have used a binomial expansion and the assumption that the Hall angle is small.

Now we rewrite Eq. (7-26) as

$$M_t = \frac{-\Delta J_i}{J_0} = \frac{-J_{xi}\omega^2\tau_i^2/2}{J_{xi}}$$

Substituting $J_{xi} = \sigma_i\mathcal{E}_x$ and summing gives

$$M_t = (eB/m)^2 <\tau^3>/<\tau> \tag{7-27}$$

where we can drop the factor of $(-\frac{1}{2})$, since the calculation is only approximate. Equation (7-27) shows that the magnetoresistance represents yet another form of scattering average. We can define, in analogy to the definitions given above, a *magnetoresistance mobility* by

$$\mu_M = \sqrt{M_t}/B = (e/m)\sqrt{<\tau^3>/<\tau>} \tag{7-28}$$

This quantity is not used very much in experimental work, but is useful as an aid in visualizing magnetoresistivity. Because of their dependence on the existence of a magnetic field, the Hall coefficient and magnetoresistivity are called *galvanomagnetic coefficients*.

7-2 The Boltzmann Transport Equation

Consider a block of semiconducting material to which an electric field and a temperature gradient have been applied. The electrons in this block are moving under the influence of both the electric and thermal forces and are not in equilibrium, since the conditions

$$\mathbf{grad}\ \bar{\mu} = 0, \qquad \mathbf{grad}\ T = 0$$

are not satisfied. Let $f(\mathbf{r}, \mathbf{p}, t)$ denote that fraction of the total number of electrons whose position coordinates lie between $\mathbf{r}$ and $\mathbf{r} + d\mathbf{r}$, and whose momentum values fall in the range $\mathbf{p}$ to $\mathbf{p} + d\mathbf{p}$ at a time t. At a later time $t + dt$, this same fraction will have its coordinates changed to $\mathbf{r} + \dot{\mathbf{r}}\,dt$ and its momenta to $\mathbf{p} + \dot{\mathbf{p}}\,dt$ (where $\dot{\mathbf{r}} = d\mathbf{r}/dt$) if the electrons do not undergo a discontinuous process such as a collision with an atom of the lattice. By using Taylor's theorem, the change in f owing to the drift motion of the electrons (i.e., the motion caused solely by the electrical and thermal gradients) is

$$f(\mathbf{r} + \dot{\mathbf{r}}dt, \mathbf{p} + \dot{\mathbf{p}}dt, t + dt) - f(\mathbf{r}, \mathbf{p}, t) = (\dot{\mathbf{r}}\cdot\nabla_r f + \dot{\mathbf{p}}\cdot\nabla_p f + \partial f/\partial t)dt \qquad (7\text{-}29)$$

where

$$\left.\begin{aligned} \nabla_r &= \mathbf{i}\,\partial/\partial x + \mathbf{j}\,\partial/\partial y + \mathbf{k}\,\partial/\partial z = \mathbf{grad} \\ \nabla_p &= \mathbf{i}\,\partial/\partial p_x + \mathbf{j}\,\partial/\partial p_y + \mathbf{k}\,\partial/\partial p_z = \mathbf{grad}_p \end{aligned}\right\} \qquad (7\text{-}30)$$

PROBLEM 7-2 Verify Eq. (7-29) by using Taylor's theorem (Appendix A1) and expanding the scalar products above.

Because of collisions, some of the electrons will have their momentum values taken out of the range specified above and others will be brought in. Let the net change be denoted by the special symbol $\partial f/\partial t)_{\text{coll}}\,dt$; equating this quantity to the right-hand side of Eq. (7-29) gives the *Boltzmann transport equation*

$$\dot{\mathbf{r}}\cdot\nabla_r f + \dot{\mathbf{p}}\cdot\nabla_p f + \partial f/\partial t = \partial f/\partial t)_{\text{coll}} \qquad (7\text{-}31)$$

Consider now what happens to $f(\mathbf{r}, \mathbf{p}, t)$ when the external gradients are removed. The material returns to equilibrium and f takes on a value f_0 which is independent of time and position.

We would now like to change the variable from the momentum $\mathbf{p}$ to the energy E. Using

$$E = \hbar\omega \qquad (1\text{-}49)$$

and the idea from Section 1-6 that the particle velocity of an electron equals its group velocity, we have

$$v_i = \frac{1}{\hbar}\frac{\partial E}{\partial k_i} \qquad (i = 1, 2, 3) \qquad (7\text{-}32)$$

from which

$$\mathbf{p} = m_e\,\mathbf{grad}_p\,E \qquad (7\text{-}33)$$

When we make this change, the fraction of electrons in the range E, $E + dE$ for equilibrium conditions is $f_0(E)$. We immediately recognize this as the Fermi distribution function of Chapter 2, since the probability that an electron will have its energy in a certain range equals the fraction lying in that range. With this in mind, we can simplify the right-hand side of Eq. (7-31) by assuming the existence of a *collision relaxation time* τ, defined as the time required for f to decay to f_0 when the applied gradients are suddenly removed. This decay depends upon the nature of the collision processes, and since little is known about them, we will assume that for a given material under a given set of conditions the decay process always requires the same length of time τ. The definition of τ can then be expressed as

$$\partial f/\partial t)_{\text{coll}} = (f - f_0)/\tau \qquad (7\text{-}34)$$

The subsequent discussion will be restricted to *steady-state* conditions, defined by

$$\partial f/\partial t = 0 \qquad (7\text{-}35)$$

Putting the last two relations into Eq. (7-31) gives

$$\dot{\mathbf{r}}\cdot\nabla_r f + \dot{\mathbf{p}}\cdot\nabla_p f = (f - f_0)/\tau \qquad (7\text{-}36)$$

PROBLEM 7-3 Show from Eq. (7-34) that the decay of f to f_0 requires a time τ.

7-3 The Solution of the Boltzmann Equation

The solution of Eq. (7-36) to be given here is based on the procedure given by Bullis (1956),[2] and is accomplished by the following steps, which correspond to the various terms in the equation reading from right to left.

(1) Assume that f represents a small deviation from the equilibrium value f_0. That is, we express f as a Taylor series, keeping only two terms, as follows:

$$f(E) = f_0(E) + a\,\partial f_0/\partial E \tag{7-37}$$

Our problem then becomes one of determining the unknown coefficient a. For convenience, we convert it into a new unknown G, by the following relation:

$$a = -\mathbf{G}\cdot\nabla_p E \tag{7-38}$$

(The reason for this definition will soon become apparent.)

(2) The term in $\Delta_p f$, from Eq. (7-37), is

$$\nabla_p f = \nabla_p f_0 + \nabla_p(a\,\partial f_0/\partial E)$$

(3) The term $\dot{\mathbf{p}}$ is, by Eqs. (1-4) and (1-14), the force on a charge carrier

$$\mathbf{F} = \dot{\mathbf{p}} = e\boldsymbol{\mathcal{E}}$$

where $\boldsymbol{\mathcal{E}}$ is the electric field and $e = +1.6 \times 10^{-19}$ coulombs. We now make the approximation that a is small enough so that the term $a\boldsymbol{\mathcal{E}}$ is negligible. It will shortly be shown that the consequence of this approximation is to guarantee the validity of Ohm's law; i.e., the current will be proportional to the electric field. We then have

$$\dot{\mathbf{p}}\cdot\nabla_p f = e\boldsymbol{\mathcal{E}}\cdot\nabla_p f_0 = e\boldsymbol{\mathcal{E}}\cdot\nabla_p E\,\partial f_0/\partial E \tag{7-39}$$

(4) The term $\nabla_r f$ becomes

$$\nabla_r f = (-\partial f_0/\partial E)[T\nabla_r(E_F/T) + (E/T)\nabla_r T] \tag{7-40}$$

PROBLEM 7-4 Verify Eq. (7-40) by utilizing

(a) the definition of f,
(b) the definition of f_0,
(c) $\nabla_r E = 0$,
(d) $\nabla_r(\partial f_0/\partial E) = \partial^2 f_0/\partial E^2\,\nabla_r E$.

---

(5) The term $\dot{\mathbf{r}}$ is written

$$\dot{\mathbf{r}} = \mathbf{v} = \nabla_p E \tag{7-41}$$

Substituting Eqs. (7-37) through (7-41) into Eq. (7-36) gives the following equation for $\mathbf{G}$:

$$-\tau[T\,\nabla_r(E_F/T) + (E/T)\,\nabla_r T] + e\tau\,\boldsymbol{\mathcal{E}} = \mathbf{G} \tag{7-42}$$

This is the desired solution.

7-4 Calculation of the Electrical and Heat Current Densities

For a material having n electrons per unit volume, all moving with velocity $\mathbf{v}$, the electrical current-density $\mathbf{J}$ is

$$\mathbf{J} = ne\mathbf{v}$$

When the velocities are not all alike,

$$J = e\int \mathbf{v}\,dn \tag{7-43}$$

where dn represents the number of electrons per unit volume whose velocities lie in the range $\mathbf{v}$, $\mathbf{v} + d\mathbf{v}$. This velocity range corresponds to a momentum-range $\mathbf{p}$, $\mathbf{p} + d\mathbf{p}$ and, by Eq. (2-63), dn can be expressed

$$V\,dn = dN = (2/h^3)f\,d\,V_p \tag{7-44}$$

where V is the physical volume of the system and dV_p is an element of volume in phase space. Substituting Eqs. (7-37), (7-41), (7-42) and (7-44) into Eq. (7-43), and utilizing the fact that the term in f_0 makes no contribution to the current, since the corresponding electrons are in equilibrium, the expression for $\mathbf{J}$ becomes

$$\mathbf{J} = (2e/h^3)\int \nabla_p E\{e\tau\boldsymbol{\mathcal{E}} - \tau[T\,\nabla_r(E_F/T)\,\nabla_r T]\}\cdot\nabla_p E\,\partial f_0/\partial E\,dV_p \tag{7-45}$$

To evaluate this integral, the following procedure is used:

(1) Use the symbol $\mathbf{A}$ to denote the quantity in the braces of Eq. (7-45), so that

$$\mathbf{J} = (2e/h^3)\int \nabla_p E(\mathbf{A}\cdot\nabla_p E)\,\partial f_0/\partial E\,dV_p \tag{7-46}$$

(2) As described in Chapter 5, some energy surfaces are ellipsoids of the form

$$E = \frac{p_x^2}{2m_x} + \frac{p_y^2}{2m_y} + \frac{p_z^2}{2m_z} \tag{5-51}$$

where E is expressed with respect to an arbitrary origin. We shall discuss the solution of the Boltzmann equation for such surfaces, initially; later, more complex surfaces will be considered.

PROBLEM 7-5 Show that the components of $\nabla_p E$ are

$$\nabla_{p_i} E = p_i/m_i, \qquad i = x, y, z \tag{7-47}$$

---

(3) In Chapter 2 we showed that there are $2/h^3$ states per unit volume of phase space, a demonstration which was based on the solution of the Schroedinger equation for a potential box considered in Chapter 1. With reference to Chapter 1, it will be seen that the transfer from an energy space to phase space was accomplished via the relation $E = p^2/2m$. Comparison of this simple relation with Eq. (5-51) shows that this is equivalent to the assumption of spherical energy surfaces. For ellipsoidal surfaces, the calculation is more complicated, but we shall assume that $2/h^3$ still specifies the number of states per unit phase space volume.

(4) The element dV of volume in phase-space can be written

$$dV_p = dS_p\, dp_n \tag{7-48}$$

where dS_p is a surface element of a constant-momentum surface and dp_n is a displacement in a direction normal to this surface. It is convenient for the applications which will be discussed later to express the integration in the normal direction in terms of energy E, and this is done by using Eq. (A-39) of Appendix A-3,

$$dE = |\nabla_p E|\, dp_n \tag{7-49}$$

so that Eq. (7-48) becomes

$$dV_p = dS_p\, dE/\nabla_p E \tag{7-50}$$

Because of the ellipsoidal energy surfaces, the calculation is simplified by introducing new coordinates.

$$\zeta_i = p_i/m_i^{1/2} \qquad i = x, y, z \tag{7-51}$$

Then the following relations hold:

$$E = (\zeta_1^2 + \zeta_2^2 + \zeta_3^2)/2 = \zeta^2/2 \tag{7-52}$$

$$\begin{aligned} |\nabla_\zeta E| &= (\zeta_1^2 + \zeta_2^2 + \zeta_3^2)^{1/2} = \zeta \\ &= (2E)^{1/2} \end{aligned} \tag{7-53}$$

$$\begin{aligned} dV_p &= (m_x m_y m_z)^{1/2}\, dV\zeta \\ &= (m_x m_y m_z)^{1/2}\, dS\zeta\ dE/|\nabla\zeta E| \end{aligned} \tag{7-54}$$

and, if we express $dS\zeta$ in spherical coordinates,

$$dS\zeta = \zeta^2 \sin\theta\, d\theta\, d\varphi \tag{7-55}$$

PROBLEM 7-6 Show that Eq. (7-46) can be expressed in component form as

$$J_i = (2e/h^3)(m_x m_y m_z m_i)^{1/2} \int \zeta_i (A_x m_x^{1/2}\zeta_x + A_y m_y^{1/2}\zeta_y + A_z m_z^{1/2}\zeta_z)(\partial f_0/\partial E)\zeta \sin\theta\, d\theta\, d\varphi \qquad (i = x, y, z) \tag{7-56}$$

(5) Consider as an example the integral involved in the A_x-term of J_x, which we denote by I_{xx}

$$\begin{aligned} I_{xx} &= m_x \int_0^{2\pi}\int_0^{\pi} \zeta_x^2 \zeta \sin\theta\, d\theta\, d\varphi \\ &= m_x \zeta^3 \int_0^{2\pi}\int_0^{\pi} (\cos\varphi \sin\theta)^2 \sin\theta\, d\theta\, d\varphi \\ &= (2E)^{3/2}\, m_x(4\pi/3) \end{aligned}$$

For the A_y-term of J_x, we find $I_{xy} = 0$, and integration of the other seven terms shows that I_{yy} and I_{zz} are identical to I_{xx}, with m_x replaced by m_y or m_z, and all other integrals vanish.

The results of steps (1) to (5) above are, therefore,

$$J_i = (2e/h^3)(m_x m_y m_z)^{1/2}\, 2^{7/2}(\pi/3) m_i \int A_i (\partial f_0/\partial E) E^{3/2}\, dE \tag{7-57}$$

Let

$$I_l = \int_0^\infty \tau E^{l+1/2}\, \partial f_0/\partial E\, dE \tag{7-58}$$

$$C_i = (2e/h^3)(m_x m_y m_z)^{1/2}\, 2^{7/2}(\pi/3) m_i$$

where the I_l are called *transport integrals*. Then

$$J_i = C_i\{[e\mathcal{E}_i - \partial E_F/\partial i + (E_F/T)\, \partial T/\partial i] I_1 - (I_2/T)\, \partial T/\partial i\} \qquad (i = x, y, z) \tag{7-59}$$

The heat current density $\mathbf{J}_Q$ was defined in Chapter 6 as the number of joules flowing across a square meter of surface per second. The equation corresponding to (7-43) is

$$\mathbf{J}_Q = \int E\mathbf{v}\, dn \tag{7-60}$$

PROBLEM 7-7 Verify that $\mathbf{J}_Q$ in Eq. (7-60) has the proper units.

PROBLEM 7-8 Show that the equation corresponding to Eq. (7-59) is

$$J_{Qi} = (C_i/e)\{[e\mathcal{E}_i - \partial E_F/\partial i + (E_F/T)\, \partial T/\partial i] I_2 - (I_3/T)\, \partial T/\partial i\} \tag{7-61}$$

We can express the electrical conductivity σ in terms of the transport integrals if we first rewrite Eq. (7-59) as follows:

$$J_i = C_i[e\mathcal{E}_i - T\,\partial/\partial T(E_F/T)\,\partial T/\partial i]\,I_1 - C_i(I_2/T)\,\partial T/\partial i \tag{7-62}$$

From Chapter 6, it is known that σ is the ratio of J_i to $\mathcal{E}_i$ when **grad** $T = 0$. Hence,

$$\sigma = C_i e I_1 \tag{7-63}$$

7-5 Solution of the Boltzmann Equation in the Presence of a Magnetic Field for Ellipsoidal Energy Surfaces

The presence of a magnetic field enormously complicates the solution of the Boltzmann equation, and we shall consider only the case **grad** $T = 0$. The force on the charged carrier is now

$$\dot{\mathbf{p}} = e[\mathcal{E} + \mathbf{v} \times \mathbf{B}] \tag{7-64}$$

and for a homogenous isothermal material we have

$$\nabla_r f = 0 \tag{7-65}$$

By using the above relations and vector identities from Appendix A2, Eq. (7-36) becomes

$$e\tau\mathcal{E}\cdot\nabla_p E - e\tau\mathbf{B}\cdot\nabla_p E \times \nabla_p a - a = 0 \tag{7-66}$$

PROBLEM 7-9 Verify Eq. (7-66).

We now wish to introduce the additional refinement that the relaxation time can be a tensor* of the form

$$\boldsymbol{\tau} = \begin{bmatrix} \tau_x & 0 & 0 \\ 0 & \tau_y & 0 \\ 0 & 0 & \tau_z \end{bmatrix}$$

As discussed in the appendix, we would then expect that Eq. (7-66) is the scalar form of the equation

$$e\boldsymbol{\tau}\mathcal{E}\cdot\nabla_p E - e\boldsymbol{\tau}\mathbf{B}\cdot\nabla_p E \times \nabla_p a - a = 0 \tag{7-67}$$

and Eq. (7-67) reduces to Eq. (7-66) when $\tau_x = \tau_y = \tau_z = \tau$.

---

*An introduction to tensor analysis covering the applications made in this chapter will be found in Appendix A5.

PROBLEM 7-10 For the first term of Eq. (7-67), which one of the following indicates the correct order of operation: $(\boldsymbol{\tau}\mathcal{E})\cdot\nabla_p E$ or $\boldsymbol{\tau}(\mathcal{E}\cdot\nabla_p E)$?

When we use Eq. (7-38), Eq. (7-67) becomes

$$e\boldsymbol{\tau}\mathcal{E}\cdot\nabla_p E + e\boldsymbol{\tau}\mathbf{B}\cdot\nabla_p E \times \nabla_p(\mathbf{G}\cdot\nabla_p E) + \mathbf{G}\cdot\nabla_p E = 0 \tag{7-68}$$

By using a vector identity, the term $\nabla_p(\mathbf{G}\cdot\nabla_p E)$ becomes

$$\nabla_p(\mathbf{G}\cdot\nabla_p E) = (\mathbf{G}\cdot\nabla_p)\nabla_p E + \mathbf{G} \times (\nabla_p \times \nabla_p E) + (\nabla_p E\cdot\nabla_p)\mathbf{G} + \nabla_p E \times (\nabla_p \times \mathbf{G}) \tag{7-69}$$

The first term on the right can be written

$$(\mathbf{G}\cdot\nabla_p)\nabla_p E = \mathbf{G}\boldsymbol{\mu}$$

where $\boldsymbol{\mu}$ is called the *reciprocal effective-mass tensor.*

PROBLEM 7-11

(a) Using Eq. (5-49), show that

$$\boldsymbol{\mu} = \begin{bmatrix} 1/m_x & 0 & 0 \\ 0 & 1/m_y & 0 \\ 0 & 0 & 1/m_z \end{bmatrix} \tag{7-70}$$

thus explaining its name.

(b) Show that the second term on the right of Eq. (7-69) is identically zero.

We must now assume that $\mathbf{G}$ is a function only of the energy E and does not explicitly depend upon the p_i (the implications of this will be discussed later). Then we can write

$$\nabla_p \times \mathbf{G} = \nabla_p E \times \partial\mathbf{G}/\partial E$$

and the fourth term on the right of Eq. (7-69) becomes

$$\nabla_p E \times (\nabla_p E \times \partial\mathbf{G}/\partial E) = \nabla_p E(\nabla_p E\cdot\partial\mathbf{G}/\partial E) - \nabla_p E\cdot\nabla_p E\,\partial\mathbf{G}/\partial E$$

The negative term of the above equation is seen to be equivalent to the third term on the right of Eq. (7-69). Hence, Eq. (7-68) now becomes

$$e\boldsymbol{\tau}\mathcal{E}\cdot\nabla_p E + e\boldsymbol{\tau}\mathbf{B}\cdot\nabla_p E \times [\mathbf{G}\boldsymbol{\mu} + \nabla_p E(\nabla_p E\cdot\partial\mathbf{G}/\partial E)] + \mathbf{G}\cdot\nabla_p E = 0$$

In evaluating the second term of this equation, we note that $e\boldsymbol{\tau}\mathbf{B}\cdot\nabla_p E \times \nabla_p E(\nabla_p E\cdot\partial\mathbf{G}/\partial E)$ vanishes, since the first operation to be performed is the one involving $\nabla_p E \times \nabla_p E$. Equating the coefficient of $\nabla_p E$ to zero then gives

$$e\boldsymbol{\tau}\boldsymbol{\mathcal{E}} + e\boldsymbol{\tau}(\mathbf{G}\boldsymbol{\mu}) \times \mathbf{B} + \mathbf{G} = 0 \tag{7-71}$$

This is the solution to the Boltzmann equation when an electric and magnetic field (but zero temperature gradient) are present in a semiconductor with ellipsoidal energy surfaces.

PROBLEM 7-12

(a) Write the vector equation [Eq. (7-71)] as three scalar equations.

(b) Solve these equations for G_x, G_y, and G_z by the use of determinants (Cramer's rule).

(c) Add the solutions to show that

$$\mathbf{G} = e\boldsymbol{\tau}[\boldsymbol{\mathcal{E}} + e\boldsymbol{\tau}\boldsymbol{\mathcal{E}}\boldsymbol{\mu} \times \mathbf{B} + e^2(\boldsymbol{\mathcal{E}}\cdot\mathbf{B})(\boldsymbol{\alpha}\mathbf{B})]/\Delta \tag{7-72}$$

where
$$\boldsymbol{\alpha} = \begin{bmatrix} (\tau_y\tau_z/m_ym_z) & 0 & 0 \\ 0 & (\tau_x\tau_z/m_xm_z) & 0 \\ 0 & 0 & (\tau_x\tau_y/m_xm_y) \end{bmatrix}$$

and the quantity $\Delta = 1 + e^2(\boldsymbol{\alpha}\mathbf{B})\cdot\mathbf{B}$ is the determinant of the coefficients of the G_i in part (b).

(d) Show that when $\boldsymbol{\mathcal{E}}$ and $\mathbf{B}$ are colinear and the energy-surfaces are spherical, then Eq. (7-72) reduces to

$$\mathbf{G} = e\boldsymbol{\tau}\boldsymbol{\mathcal{E}}$$

which is the value it would have for $B = 0$.

The electrical current density is, by Eqs. (7-42) and (7-45),

$$\mathbf{J} = (2e/h^3) \int \nabla_p E(\mathbf{G}\cdot\nabla_p E)\, \partial f_0/\partial E\, dV_p \tag{7-73}$$

where

(a) $$\mathbf{G}\cdot\nabla_p E = e\left[\sum_{ij} \frac{\tau_i}{m_j}\, \mathcal{E}_i p_j\, \delta_{ij} + e\sum_{ijk} \frac{\tau_i\tau_j}{m_im_j}\, \mathcal{E}_i p_j B_k \epsilon_{ijk} \right.$$
$$\left. + e^2\xi(\mathbf{B}\cdot\mathbf{p}) \sum_{ij} \mathcal{E}_i B_j \delta_{ij} \right]/\Delta \tag{7-74}$$

(b) δ_{ij} is the Kronecker delta (see appendix).

(c) $\xi = \tau_x\tau_y\tau_z/m_xm_ym_z$

(d) ϵ_{ijk} is the permutation symbol, defined in the appendix by Eq. (A-27).

The conductivity tensor will connect $\mathbf{J}$ and $\boldsymbol{\mathcal{E}}$ in accordance with the relation

$$\mathbf{J} = \boldsymbol{\sigma}\boldsymbol{\mathcal{E}}, \quad \text{or} \quad J_g = \sum_i \sigma_{gi}\mathcal{E}_i \tag{7-75}$$

Putting Eq. (7-74) into Eq. (7-73) and comparing with Eq. (7-75) shows that

$$\sigma_{gi} = \sum_{ijk} (2e^2/h^3m_gm_i) \int [\tau_i p_g p_i + (e/m_j)\tau_i\tau_j p_g p_j B_k \epsilon_{ijk}$$
$$+ (e^2 m_i \xi \mathbf{B}\cdot\mathbf{p}\; p_g B_i)]\, \partial f_0/\partial E\, dV_p/\Delta \tag{7-76}$$

As in Section 7-4, the integration is performed by integrating over a constant energy surface and over the entire range of energies. The result is then a generalization of Eq. (7-57).

$$\sigma_{gi} = \frac{2^{9/2}e^2\sqrt{m_xm_ym_z}\,\pi}{3h^3}\sum_{ijk}\int_0^\infty \left(\frac{\tau_i\,\delta_{gi}}{\sqrt{m_gm_i}}\right.$$
$$\left. + \frac{e\tau_i\tau_j\,\delta_{gj}\,\epsilon_{ijk}B_k}{m_i\sqrt{m_gm_j}} + e^2B_gB_i\right)E^{3/2}\,\frac{\partial f_0}{\partial E}\,dE/\Delta \tag{7-77}$$

Since the ellipsoidal surfaces are ellipsoids of revolution, it is convenient at this point to introduce the transverse and longitudinal effective mass m_t and m_l of Section 5-9, plus the corresponding relaxation times τ_t and τ_l. We then define an *anisotropy ratio* K, by

$$K = \tau_t m_l/\tau_l m_t \tag{7-78}$$

where $m_t = m_1 = m_2$, $\quad m_l = m_3$

It is also convenient to assume that τ_t and τ_l have the same functional dependence on energy; that is, let

$$\tau_t = \alpha_t\tau(E), \qquad \tau_l = \alpha_l\tau(E) \tag{7-79}$$

Then Eq. (7-77) becomes

$$\boldsymbol{\sigma} = g\begin{bmatrix} K(I_{11} + v^2KB_x^2I_{13}) & vK^2(-B_zI_{12} + vB_xB_yI_{13}) & vK(B_yI_{12} + vKB_xB_zI_{13}) \\ vK^2(B_zI_{12} + vB_xB_yI_{13}) & K(I_{11} + v^2KB_y^2I_{13}) & vK(B_xI_{12} + vKB_yB_zI_{13}) \\ vK(-B_yI_{12} + vKB_xB_zI_{13}) & vK(-B_xI_{12} + vB_yB_zI_{13}) & (I_{11} + v^2K^2B_z^2I_{13}) \end{bmatrix} \tag{7-80}$$

where
$$g = 2^{9/2}\pi e^2 m_t\alpha_l/3h^3m_l^{1/2} \tag{7-81}$$

$$\left.\begin{aligned} v &= e\alpha_l/m_l \\ K &= \alpha_t m_l/m_t\alpha_l \end{aligned}\right\} \tag{7-82}$$

and the transport integrals of Eq. (7-58) are now

$$I_{1n} = \int_0^\infty \tau^n(E)\, E^{3/2}\, (\partial f_0/\partial E)\, dE/\Delta \tag{7-83}$$

where
$$\Delta = 1 + v^2\tau^2 K(B_x^2 + B_y^2 + KB_z^2) \tag{7-84}$$

Note that $l = 1$; i.e., $E^{l+1/2} = \mathrm{E}^{3/2}$, for all three of the integrals in Eq. (7-83).

7-6 The Galvanomagnetic Coefficients

When an electric current is flowing in a material whose conductivity is a tensor $\boldsymbol{\sigma}$, then $\mathbf{J}$ and $\boldsymbol{\mathcal{E}}$ are not colinear. We have shown that this will be the case for a semiconductor in a magnetic field $\mathbf{B}$, and when $\mathbf{B} = 0$ in Eq. (7-80), we note that $\boldsymbol{\sigma}$ reduces to a scalar. It is convenient to resolve $\boldsymbol{\mathcal{E}}$ into three orthogonal components, with directions as follows:

(1) parallel to $\mathbf{J}$.
(2) normal to the plane of $\mathbf{J}$ and $\mathbf{B}$.
(3) normal to $\mathbf{J}$ and lying in the $(\mathbf{J}, \mathbf{B})$-plane.

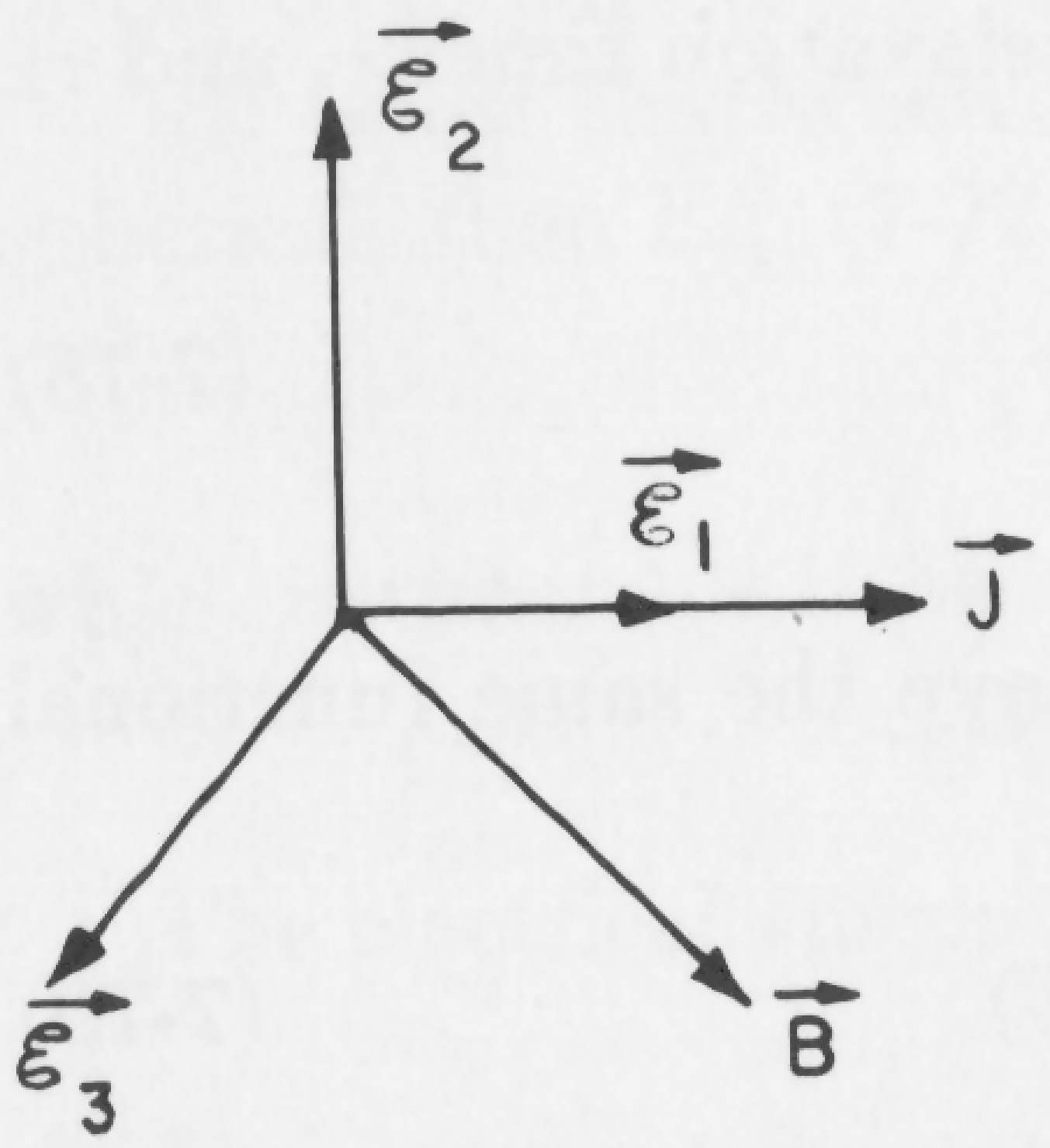


FIG. 7-2. The components of the electric field in a semiconductor.

These directions are illustrated in Fig. 7-2; for convenience, we shall label the axes 1, 2, and 3 to correspond to the components of $\boldsymbol{\mathcal{E}}$.

Vectorially, these directions may be represented by

$$(1)\ \mathbf{J}, \qquad (2)\ \mathbf{J} \times \mathbf{B}, \qquad (3)\ (\mathbf{J} \times \mathbf{B}) \times \mathbf{J}$$

A galvanomagnetic coefficient is associated with each direction in the following manner:

(1) The magnetoresistance coefficient

$$M = [(\boldsymbol{\mathcal{E}}\cdot\mathbf{J})/(\boldsymbol{\mathcal{E}}\cdot\mathbf{J})_{B=0}] - 1 \tag{7-85}$$

(2) The Hall coefficient

$$R = (\boldsymbol{\mathcal{E}}\cdot\mathbf{J} \times \mathbf{B})/(\mathbf{J} \times \mathbf{B})^2 \tag{7-86}$$

(3) The *planar Hall coefficient*

$$P = \boldsymbol{\mathcal{E}}\cdot(\mathbf{J} \times \mathbf{B}) \times \mathbf{J}/(\mathbf{J} \times \mathbf{B})^2(\mathbf{J}\cdot\mathbf{B}) \tag{7-87}$$

Let us apply these definitions to the case of $\mathbf{J}$ and $\mathbf{B}$ parallel. Choosing this direction as the 3-axis, we obtain

$$\mathbf{J} = \mathbf{i}_3 J, \qquad \mathbf{B} = \mathbf{i}_3 B \tag{7-88}$$

where $\mathbf{i}_3$ is a unit vector along the 3-axis. Under these conditions, M is called the *longitudinal magnetoresistance* M_l.

PROBLEM 7-13

(a) Use Eq. (7-88) to show that Eq. (7-80) reduces to the form

$$\boldsymbol{\sigma} = \begin{bmatrix} \sigma_{11} & -\sigma_{12} & 0 \\ \sigma_{12} & \sigma_{11} & 0 \\ 0 & 0 & \sigma_{33} \end{bmatrix} \tag{7-89}$$

(b) Show that the corresponding *resistivity tensor*, defined by $\boldsymbol{\mathcal{E}} = \boldsymbol{\rho}\mathbf{J}$, will have the form

$$\boldsymbol{\rho} = \begin{bmatrix} \rho_{11} & \rho_{12} & 0 \\ -\rho_{12} & \rho_{11} & 0 \\ 0 & 0 & \rho_{33} \end{bmatrix} \tag{7-90}$$

and express each ρ_{ij} in terms of the σ_{ij}.

(c) Using Eq. (7-90), show that

$$\boldsymbol{\mathcal{E}}\cdot\mathbf{J} = \rho_{33} J^2 \tag{7-91}$$

(d) Show that

$$M_l + 1 = \rho_{33}/\rho_0 \tag{7-92}$$

where
$$\rho_0 = \rho_{11})_{B=0} = \rho_{33})_{B=0}$$

or
$$M_l = (\rho_{33} - \rho_0)/\rho_0 \tag{7-93}$$

which is of the same form as Eq. (7-26). Show also that M_l is the ratio of the electrical power transmitted with and without a magnetic field present.

PROBLEM 7-14

(a) Let $\mathbf{J} = \mathbf{i}_1 J$ and $\mathbf{B} = \mathbf{i}_3 B$. Show that

$$\boldsymbol{\mathcal{E}} = (\rho_{11}\mathbf{i} - \rho_{12}\mathbf{j})J \tag{7-94}$$

(b) Show that the transverse magnetoresistance M_t is given by

$$M_t + 1 = \rho_{11}/\rho_0 \tag{7-95}$$

(c) Show that the Hall coefficient is

$$R = \rho_{12}/B \tag{7-96}$$

and that the planar Hall coefficient is

$$P = 0$$

Show also that Eq. (7-86) is equivalent to Eq. (7-12).

PROBLEM 7-15 Let $\mathbf{B}$ lie along the 3-axis and $\mathbf{J}$ lie in the (2, 3)-plane, at a 45-degree angle to each axis. Show that

(a) $\boldsymbol{\mathcal{E}} = (\rho_{12}\mathbf{i} + \rho_{22}\mathbf{j} + \rho_{33}\mathbf{k})J/2$

(b) $M_{45°} + 1 = (\rho_{11} + \rho_{33})/2\rho_0$

(c) $R_{45°} = \rho_{12}/B$

(d) $P_{45°} = (\rho_{33} - \rho_{11})/B^2$

7-7 Interpretation of Experimental Data for N-Type Germanium and Silicon

In the analysis given in the last two sections, the galvanomagnetic coefficients were calculated in terms of the coefficients of $\boldsymbol{\sigma}$ or $\boldsymbol{\rho}$ for semiconductors with ellipsoidal energy surfaces, and these, in turn, involve the transport integrals I_{1n} of Eq. (7-84). We are now ready to compute explicit formulas for these coefficients that can be compared with the results of experiments on N-type germanium and silicon.

To do this, we must introduce the following two factors into the calculation:

(1) The arrangement of the ellipsoids in the conduction band of germanium and silicon.
(2) The dependence of the relaxation time τ on the energy E.

For the first item, it was shown in Section 5-9 that the conduction band for germanium has eight equivalent minima located at the ends of the cube diagonals; these correspond to half-ellipsoids, making a total of four complete ones. For silicon, Problem 5-44 indicates six ellipsoids part way along the cube axes. Let us treat germanium first, since the ellipsoid arrangement is more involved.

For a many-valley semiconductor, each ellipsoid corresponds to available energies for electrons excited from the valence band; in order to calculate the conductivity, we must add the contributions from all the ellipsoids. This introduces a geometrical problem, because the calculation so far has involved only a single ellipsoid whose axes lie along the coordinate axes. Figure 7-3 illustrates the relation between a typical ellipsoid, which we choose as the one lying along the $<\bar{1}\bar{1}1>$ axis, and the axis of the crystal. The unit vectors for the coordinate system of the ellipsoid are denoted by $\mathbf{i}$, $\mathbf{j}$, $\mathbf{k}$; those for the crystal are $\mathbf{I}$, $\mathbf{J}$, $\mathbf{K}$. Since the ellipsoid is symmetric about $\mathbf{k}$, the orientation of $\mathbf{i}$ and $\mathbf{j}$ is arbitrary, and we locate them to make the angle between $\mathbf{i}$ and $\mathbf{I}$ equal that between $\mathbf{j}$ and $\mathbf{J}$. The relative direction cosines of the two systems can be shown as lengths a, b, and c in the figure, since we are projecting unit vectors. The transformation matrix* is then

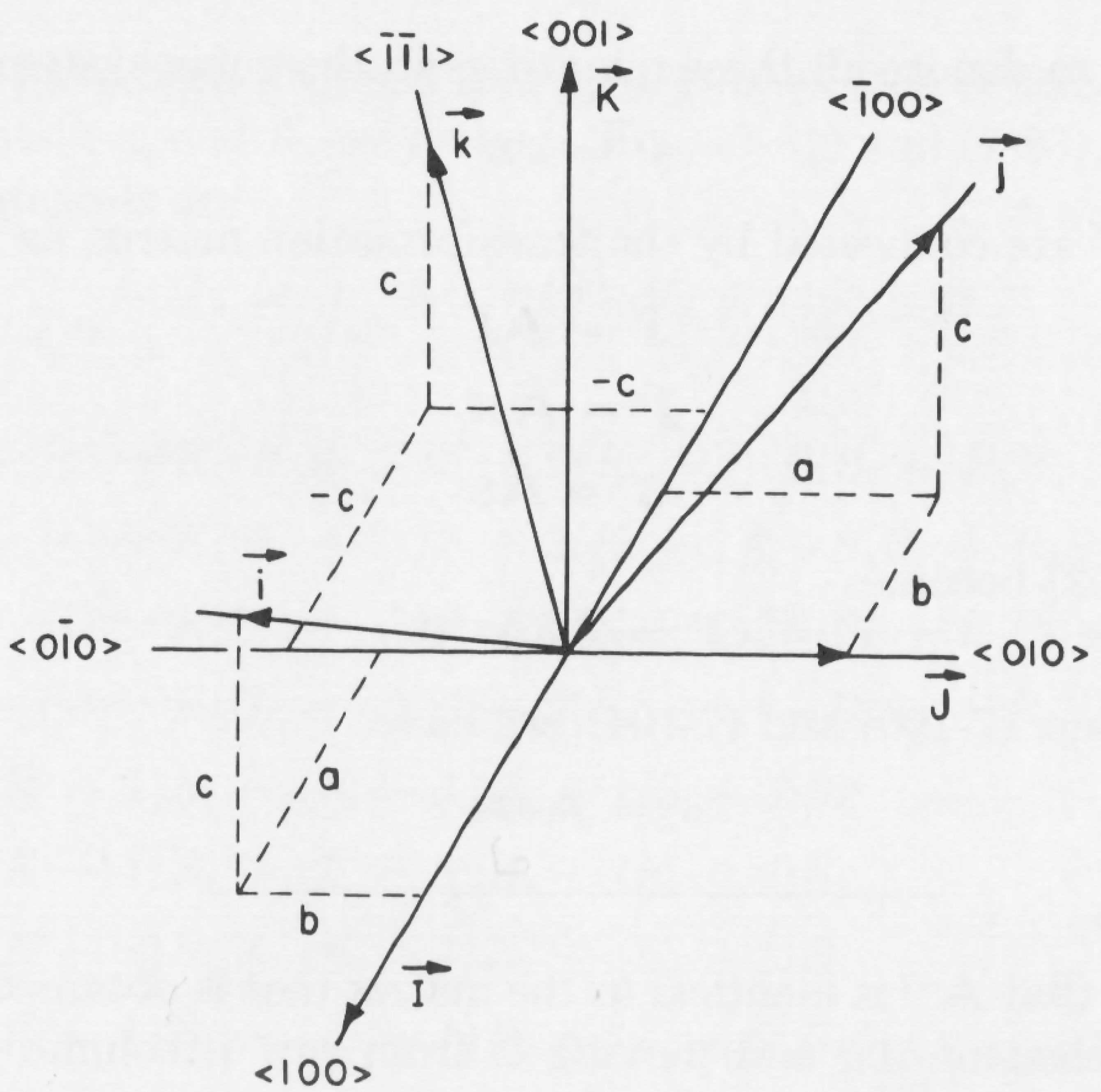


FIG. 7-3. Transformation from ellipsoid axes to crystal axes in N-type germanium. (From W. M. Bullis, Ph. D. Thesis, M. I. T., 1956, with permission.)

$$\mathbf{A} = \begin{array}{c} \\ \mathbf{i} \\ \mathbf{j} \\ \mathbf{k} \end{array}\!\!\begin{array}{c} \begin{array}{ccc} \mathbf{I} & \mathbf{J} & \mathbf{K} \end{array} \\ \begin{bmatrix} a & b & c \\ b & a & c \\ -c & -c & c \end{bmatrix} \end{array} \tag{7-97}$$

PROBLEM 7-16

(a) Find the transformation matrix for the ellipsoid along the $< 1\bar{1}1 >$ direction.

(b) Show that

$$\begin{aligned} a &= (1 + \sqrt{3})/2\sqrt{3} \\ b &= (1 - \sqrt{3})/2\sqrt{3} \\ c &= 1/\sqrt{3} \end{aligned} \tag{7-98}$$

The conductivity tensor $\boldsymbol{\sigma}$ of Eq. (7-80) is related to $\mathbf{J}$ and $\boldsymbol{\mathcal{E}}$ in the ellipsoidal system by

$$\mathbf{J} = \boldsymbol{\sigma}\boldsymbol{\mathcal{E}} \tag{7-99}$$

---

*See Appendix A4 for a discussion of matrix algebra and coordinate transformations.

Using primes to denote all these quantities in the cubic system, then

$$\mathbf{J}' = \boldsymbol{\sigma}'\boldsymbol{\mathcal{E}}' \tag{7-100}$$

Now $\mathbf{J}$ and $\mathbf{J}'$ are connected by the transformation matrix, as follows:

$$\mathbf{J}' = \mathbf{A}\mathbf{J} \tag{7-101}$$

so that

$$\mathbf{J}' = \mathbf{A}\boldsymbol{\sigma}\boldsymbol{\mathcal{E}} \tag{7-102}$$

But

$$\boldsymbol{\mathcal{E}}' = \mathbf{A}\boldsymbol{\mathcal{E}} \tag{7-103}$$

and Eq. (7-102) becomes

$$\mathbf{J}' = \mathbf{A}\boldsymbol{\sigma}\mathbf{A}^{-1}\boldsymbol{\mathcal{E}}' \tag{7-104}$$

Comparing Eqs. (7-100) and (7-104), we have

$$\boldsymbol{\sigma}' = \mathbf{A}\boldsymbol{\sigma}\mathbf{A}^{-1} \tag{7-105}$$

PROBLEM 7-17

(a) Show that $\mathbf{A}^{-1}$ is identical to the matrix that is obtained from $\mathbf{A}$ by taking each element A_{ij} and moving it from row i, column j, to row j, column i.

(b) Use Eq. (7-105) and part (a) of this problem to show that

$$\begin{aligned}\sigma'_{11} = {} & \sigma_{11}A^2_{11} + \sigma_{22}A^2_{21} + \sigma_{33}A^2_{31} + (\sigma_{12} + \sigma_{21})A_{11}A_{21} \\ & + (\sigma_{13} + \sigma_{31})A_{11}A_{31} + (\sigma_{23} + \sigma_{32})A_{31}A_{21}\end{aligned} \tag{7-106}$$

(c) From Eqs. (7-80) and (7-97), show that Eq. (7-106) gives

$$\sigma'_{11} = (g/3)(2K + 1)I_{11} + gv^2K^2I_{13}(aB_x + bB_y - cB_z)^2 \tag{7-107}$$

(d) Show that when $\mathbf{B}$ is expressed in the cubic system

$$B'_x = aB_x + bB_y - cB_z \tag{7-108}$$

(e) Show that Δ of Eq. (7-84) in the cubic system is

$$\Delta_1 = 1 + (v^2\tau^2/3)K[(K + 2)B^2 + 2(K - 1)(B'_xB'_y - B'_yB'_z - B'_zB'_x)] \tag{7-109}$$

where the notation Δ_1 refers to the ellipsoid in the $< \bar{1}\bar{1}1 >$ direction.

Making similar transformations for the other three ellipsoids and adding the four contributions, we find that the elements of the conductivity tensor are

$$\begin{aligned}\sigma_{11} &= g\sum_{j=1}^{4} [(2K + 1)(I^j_{11}/3) + v^2B^2_1K^2I^j_{13}] \\ \sigma_{12} &= g\sum_{j=1}^{4} [(K - 1)(-1)^j(I^j_{11}/3) \\ &\qquad - (v/3)Kq_jI^j_{12} + v^2B_1B_2K^2I^j_{13}]\end{aligned} \tag{7-110}$$

The primes have been dropped and x, y, z have been replaced by 1, 2, 3. The constants v, g, and K are given in Eqs. (7-81) and (7-83), and the four transport integrals are

$$I^j_{1n} = \int_0^\infty [\tau(\epsilon)]^n E^{3/2}(\partial f_0/\partial\epsilon)(1/\Delta_j)\, dE \tag{7-111}$$

where

$$\begin{aligned}\Delta_1 &= 1 + (\tfrac{1}{3})v^2\tau^2K[(K + 2)B^2 + 2(K - 1)(B_1B_2 - B_2B_3 - B_3B_1)] \\ \Delta_2 &= 1 + (\tfrac{1}{3})v^2\tau^2K[(K + 2)B^2 + 2(K - 1)(-B_1B_2 + B_2B_3 - B_3B_1)] \\ \Delta_3 &= 1 + (\tfrac{1}{3})v^2\tau^2K[(K + 2)B^2 + 2(K - 1)(B_1B_2 + B_2B_3 + B_3B_1)] \\ \Delta_4 &= 1 + (\tfrac{1}{3})v^2\tau^2K[(K + 2)B^2 + 2(K - 1)(-B_1B_2 - B_2B_3 + B_3B_1)] \\ q_1 &= -(K - 1)B_1 - (K - 1)B_2 + (K + 2)B_3 \\ q_2 &= -(K - 1)B_1 + (K - 1)B_2 + (K + 2)B_3 \\ q_3 &= +(K - 1)B_1 + (K - 1)B_2 + (K + 2)B_3 \\ q_4 &= +(K - 1)B_1 - (K - 1)B_2 + (K + 2)B_3\end{aligned}$$

The other components of $\boldsymbol{\sigma}$ can be found by cyclic permutation of the indices on the B_i (for example, σ_{22} comes from σ_{11} by replacing every B_1 with B_2).

PROBLEM 7-18

(a) Show that

$$\sigma_{21}(\mathbf{B}) = \sigma_{12}(-\mathbf{B}) \tag{7-112}$$

(b) Write Eq. (7-99) in component form and show that Eq. (7-112) is an example of the Onsager relations of Chapter 6 under the condition $\mathbf{grad}\ T = 0$.

Although these equations are complicated, they simplify tremendously when $\mathbf{B}$ is chosen along special directions.

PROBLEM 7-19 Let $\mathbf{B}$ lie along the 3-axis, so that it reduces to $\mathbf{i}_3B$.

(a) Show that

$$\Delta_1 = \Delta_2 = \Delta_3 = \Delta_4 = 1 + v^2\tau^2K(K + 2)B^2/3$$

and

$$q_1 = q_2 = q_3 = q_4 = (K + 2)B$$

(b) Show that the only non-vanishing σ_{ij} become

$$\left.\begin{aligned}\sigma_{11} &= \sigma_{22} = 4g(2K + 1)I_{11}/3 \\ \sigma_{33} &= [4g(2K + 1)/3][I_{11} + v^2K^2B^2I_{13}] \\ \sigma_{12} &= -\sigma_{21} = -4g(v/3)\,K(K + 2)\,BI_{12}\end{aligned}\right\} \tag{7-113}$$

(c) Show that the corresponding elements of $\boldsymbol{\varrho}$ are

$$\rho_{11} = \sigma_{11}/[\sigma_{11}^2 + \sigma_{21}^2] = \rho_{22}$$
$$\rho_{12} = \sigma_{21}/[\sigma_{11}^2 + \sigma_{21}^2] = -\rho_{21} \tag{7-114}$$
$$\rho_{33} = 1/\sigma_{33}$$

(d) Using the results of Problem 7-13 and 7-14, express M_l, M_t, and R in terms of the I_{1n}.

The second item mentioned at the beginning of this section is the knowledge of how $\tau(E)$ depends on E. The relaxation time τ indicates the probability that an electron will collide with some disturbance in the lattice, generally a vibrating atom or an ionized impurity. After this collision, the electron has a different energy and the number of states available for it to go into is proportional to $E^{1/2}$, in accordance with Eq. (1-88). Since the quantity $1/\tau$ can be regarded as the probability of scattering, we have

$$\tau = lE^{-1/2} \tag{7-115}$$

where l is a constant of proportionality.

PROBLEM 7-20 Show that l, called the *mean free path*, has the dimensions of length and can be interpreted as the distance travelled between collisions.

In the next chapter, we shall consider the theory of scattering and show that Eq. (7-115) applies only for collisions with the lattice vibrations; for ionized impurities, it will be seen that the connection between τ and E is

$$\tau = lE^{3/2} \tag{7-116}$$

so that l no longer is a length. Deferring the question of the nature of scattering, we shall assume that

$$\tau = lE^{\lambda} \tag{7-117}$$

where $\lambda = -\frac{1}{2}$ gives Eq. (7-115) and $\lambda = \frac{3}{2}$ gives Eq. (7-116).

For reasonably pure semiconductors, from Section 3-3, we can use the Boltzmann approximation.

$$f_0 = e^{-(E-E_F)/kT} \tag{7-118}$$

If we let

$$\chi = E/kT \tag{7-119}$$

the transport integral [Eq. (7-111)] becomes

$$I_{1\,n}^{\,j} = C \int \tau^n \chi^{3/2} e^{-\chi}\, d\chi/\Delta_j \tag{7-120}$$

where

$$C = -n\, h^3\, e^{E_F/2kT}/(2\pi)^{3/2} m_t m_l^{1/2} \tag{7-121}$$

PROBLEM 7-21 Using the results of Section 3-3, verify Eqs. (7-120) and (7-121).

The quantity Δ_j in the denominator of Eq. (7-120) can be expanded by the binomial theorem; for low fields, it becomes

$$\Delta^{-1} = 1 - \beta_j^2\tau^2 + \beta^4\tau^4 - \cdots \tag{7-122}$$

where the β_j are the various coefficients of the τ^2 in Eq. (7-111) and where we have dropped the subscript j on Δ. Equation (7-120) is now

$$I = C[<\tau^n> - \beta_j^2 <\tau^{n+2}> + \cdots] \tag{7-123}$$

where

$$<\tau^n> = \int_0^{\infty} \tau^n \chi^{3/2} e^{-\chi}\, d\chi \tag{7-124}$$

PROBLEM 7-22

(a) Show that when $\mathbf{B} = \mathbf{i}_3 B$, Eq. (7-123) gives

$$I_{11} = C\,[<\tau> - v^2K(K+2)B^2 <\tau^3> /3] \tag{7-125}$$

(b) Find similar expressions for I_{12} and I_{13}.

(c) Use the results of Problem 7-19(d) to show that the low-field Hall coefficient is

$$\begin{aligned} R_{\text{low}} &= \frac{3vK(K+2)I_2}{4g[(2K+1)^2 I_1^2 + v^2K^2(K+2)^2B^2I_2^2]} \\ &= \left\{<\tau^2> + B^2\left[\frac{2v^2K(K+2) <\tau^2> <\tau^3>}{3<\tau>}\right.\right. \\ &\quad \left.\left. - v^2K(K+2) <\tau^4> - \frac{v^2K^2(K+2)^2 <\tau^2>^3}{(2K+1)^2 <\tau^2>}\right]\right\} \Big/ C \end{aligned} \tag{7-126}$$

to terms in B^2.

(d) Show that in the limit of $B = 0$, Eq. (7-126) reduces to

$$R_0 = \frac{3vK(K+2) <\tau^2>}{4gC(2K+1) <\tau>^2} \tag{7-127}$$

Compare this with the simple expression for the Hall coefficient obtained at the beginning of this chapter, using Eqs. (7-81), (7-82), and (7-121) to evaluate the constants.

(e) Expand in powers of $1/\beta_j^2\tau^2$ for the high-field case and show that

$$R_{\text{high}} = (1/ne)(1 + A/B^2) \tag{7-128}$$

where A is a term involving K and the $< \tau^n >$. What is the physical significance of Eq. (7-128) in the limit of $B = \infty$? Show that

$$\mu_H/\mu = R_0/R_\infty = 3K(K+2) < \tau^0 > < \tau^2 > /(2K+1)^2 < \tau >^2 \tag{7-129}$$

where

$$< \tau^0 > = 3\sqrt{\pi}/4$$

$$R_\infty = 1/ne$$

The assumption of Eq. (7-117) enables us to make numerical calculations of the galvanomagnetic coefficients by introducing the *gamma function* $\Gamma(z)$, which is defined as

$$\Gamma(z) = \int_0^\infty e^{-t} t^{z-1}\, dt \tag{7-130}$$

and for which tables are available.

PROBLEM 7-23

(a) Show that

$$< \tau^n > = l^n(kT)^{n\lambda} C\Gamma(n\lambda + 2.5) \tag{7-131}$$

(b) Show that

$$< \tau^0 > < \tau^2 > / < \tau >^2 = 3\pi/8 \quad \text{for} \quad \lambda = -\tfrac{1}{2}$$

$$< \tau^0 > < \tau^2 > / < \tau >^2 = \frac{315\pi}{512} \quad \text{for} \quad \lambda = \tfrac{3}{2}$$

(c) Evaluate R_0 for spherical energy surfaces.

PROBLEM 7-24 Using a procedure similar to that above, show that

$$M_l = \frac{-e^2B^2}{m_t^2} \frac{< \tau_t^3 >}{< \tau_t >} \tag{7-132}$$

where

$$\tau_t = \alpha_t \tau$$

and compare with the calculations in Section 7-1.

PROBLEM 7-25 Defining the Hall mobility μ_H as

$$\mu_H = R_0\sigma_0$$

in accordance with Eq. (7-23), show that

$$\mu_H = \frac{vK(K+2) < \tau^2 >}{(2K+1) < \tau >} \tag{7-133}$$

PROBLEM 7-26

(a) Show that

$$M_l = (\mu_H B)^2 \frac{2(K-1)^2(2K+1) < \tau > < \tau^3 >}{3K(K+2)^2 < \tau^2 >^2}$$

(b) What is the value of M_l for spherical energy surfaces, and what is the significance of this?

(c) Show that

$$M_t = (\mu_H B)^2 \left[\frac{(2K+1)^2 < \tau > < \tau^3 >}{3K(K+2) < \tau^2 >^2} - 1 \right]$$

The variation of mobility with temperature is an important property of a semiconductor, since it determines many of the characteristics of semiconductor devices. From Eqs. (7-131) and (7-133), we see that the Hall mobility, for example, will depend on temperature according to

$$\mu_H = CT^\lambda \tag{7-134}$$

where C is a constant and $\lambda = \frac{1}{2}$ or $-\frac{3}{2}$ for the types of scattering we have been considering. Bullis and Krag (1956)[3] show that when the Hall coefficient is measured as a function of magnetic field at the temperature of liquid nitrogen (77°K), we can disregard the temperature-dependence of mobility, or the energy-dependence of the relaxation time, and treat τ as a constant. This further simplifies expressions such as Eq. (7-129), since we now have $< \tau^2 > = < \tau >^2$.

PROBLEM 7-27 Using Eqs. (7-126) and (7-133), and the condition that τ is constant, show that

$$\frac{R}{R_\infty} = \frac{3K(K+2)}{(2K+1)^2} \frac{m}{1 + z^2b^2} \tag{7-135}$$

where

$$m = 1 + \frac{(K+2)b^2}{3K} \tag{7-136}$$

$$z = \frac{K+2}{2K+1} \tag{7-137}$$

$$b = \frac{2(2K+1)}{\sqrt{3\pi K(K+2)}} \mu_H B \tag{7-138}$$

Plots of the Hall coefficient vs magnetic field in terms of R/R_H and b are shown in Fig. 7-4, and experimental points of Bullis and Krag are shown for comparison. One curve is calculated from Eq. (7-135), and the other represents the relation between R and B when the magnetic field lies at a 45-deg angle to two of the cubic axes and J along one of these axes. The equation corresponding to Eq. (7-135) can be obtained in the same manner, starting with the expressions given in Problem 7-15. The curves for Fig. 7-4 are plotted by using $K = 19$, showing how the anisotropy factor can be obtained by experimental fitting. The value of τ was determined by matching the minima on the B-axis of the lower theoretical curve and the corresponding experimental one, and it was found that $\tau = 2.4 \times 10^{-12}$ sec.

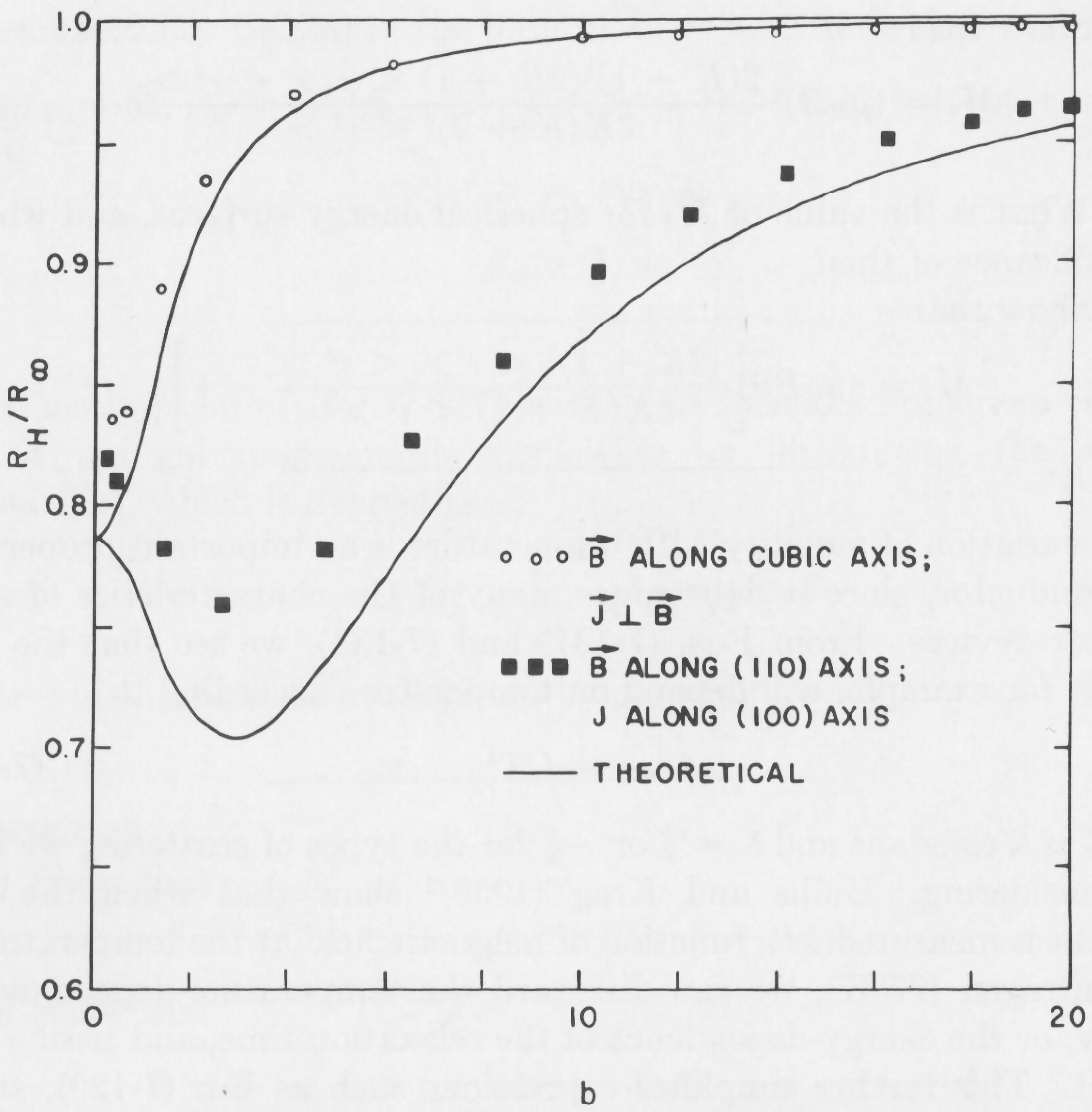


FIG. 7-4. Hall coefficient as a function of magnetic field for N-type germanium. (From W. M. Bullis and W. E. Krag, Phys. Rev. **101,** 580, 1956, with permission.)

A similar study of N-type silicon was made by Krag (1960)[4], who found that $K = 3.2$ and τ = constant gave a good fit to the experimental data (Fig. 7-5). The calculations for silicon are similar to those for germanium, but are simpler because the six energy ellipsoids lie along cubic axes, and it is not necessary to make coordinate transformations. Although the interpretation of both Figs. 7-4 and 7-5 is based on an energy-independent τ, we shall see in the next chapter the importance of the relation given by Eq. (7-117).

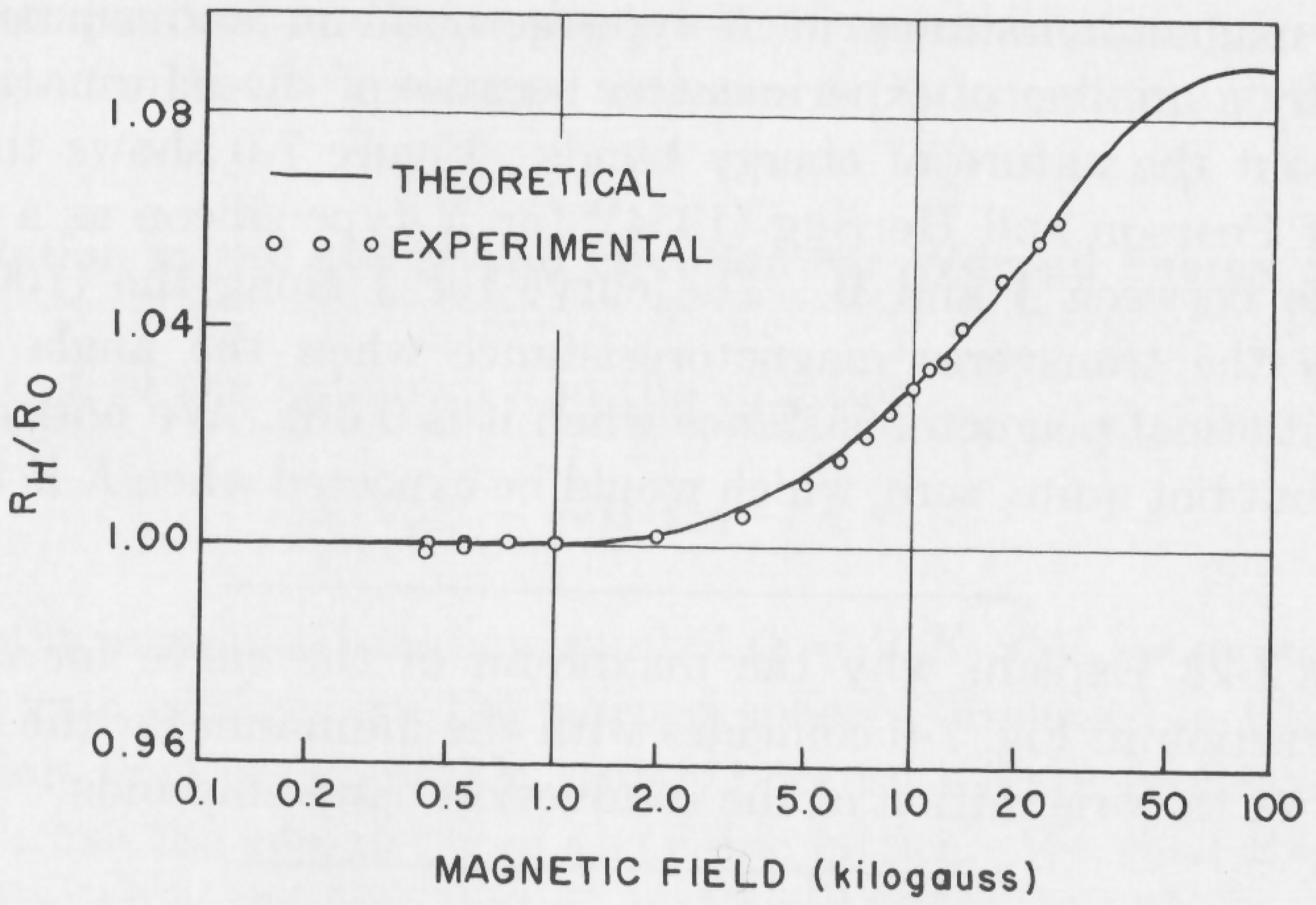


FIG. 7-5. Hall coefficient as a function of magnetic field for N-type silicon. (From W. E. Krag, Phys. Rev. **118,** 435, 1960, with permission.)

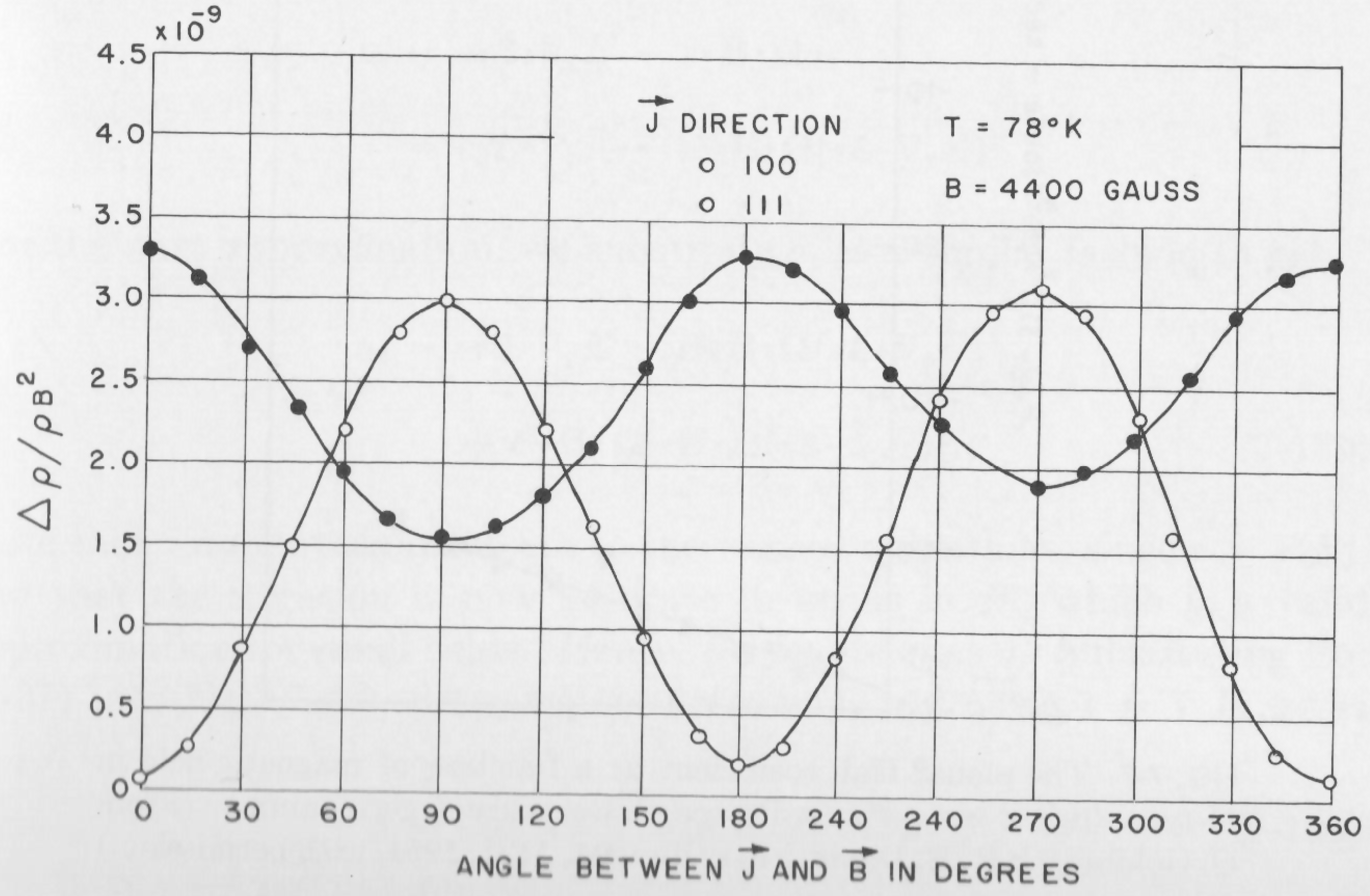


FIG. 7-6. The dependence of magnetoresistance on orientation in N-type silicon. (After G. L. Pearson and C. Herring, Physica **20,** 975, 1954, with permission.)

The magnetoresistance in N-type germanium and silicon has been studied by a number of experimenters because of the information which it gives about the nature of energy bands. Figure 7-6 shows the measurements of Pearson and Herring (1954)[5] for N-type silicon as a function of the angle between $\mathbf{J}$ and $\mathbf{B}$. The curve for $\mathbf{J}$ along the (100) direction indicates the transverse magnetoresistance when the angle is 90° and the longitudinal magnetoresistance when it is 0 deg. We notice that M_l is almost, but not quite, zero, which would be expected when K is fairly small.

PROBLEM 7-28 Explain why the maximum of the curve for J along the (111) direction in Fig. 7-6 coincides with the minimum for the other curve in terms of the orientation of the conduction band ellipsoids.

The planar Hall coefficient was defined and first measured by Goldberg and Davis (1954);[6] the dependence of $P_{45°}$ for an N-type and a P-type germanium sample as a function of B^2 is shown in Fig. 7-7. By using the expression given in Problem 7-15(d), it can be shown, as above, that $P_{45°}/B^2$ is independent of B^2 for the low-field approximation. Hence, the non-zero slope for the P-type curve indicates a B^4 dependence in this case.

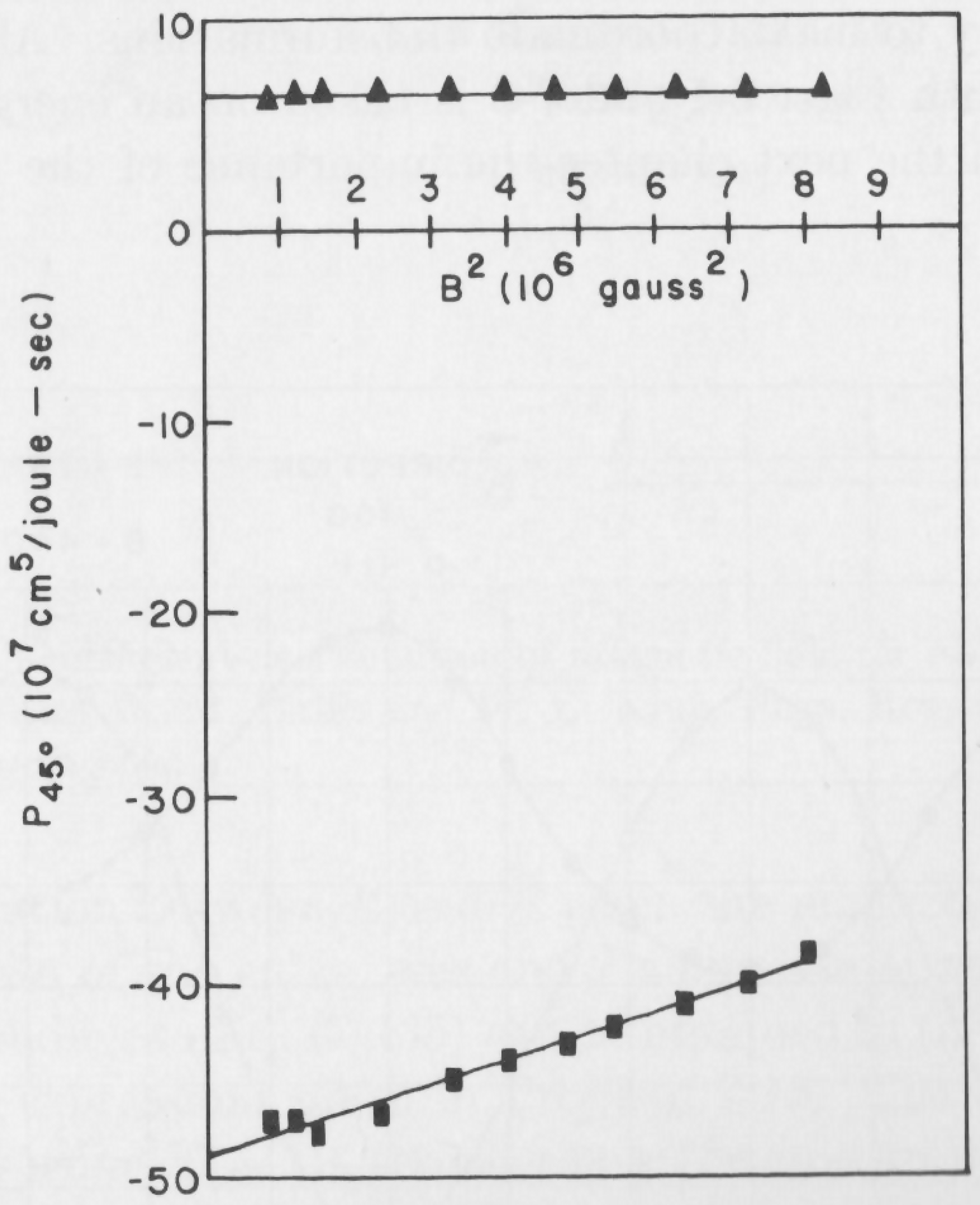


FIG. 7-7. The planar Hall coefficient as a function of magnetic field in N-type (upper curve) and P-type (lower curve) germanium. (From C. Goldberg & R. E. Davis, Phys. Rev. **94,** 1121, 1954, with permission.)

7-8 The Solution to the Boltzmann Equation for Warped Energy Surfaces

The calculation of the unknown a in the equation

$$e\tau\boldsymbol{\mathcal{E}}\cdot\nabla_p E - e\tau\mathbf{B}\cdot\boldsymbol{\Omega}a - a = 0 \tag{7-66}$$

(where we have introduced the new symbol $\boldsymbol{\Omega} = \nabla_p E \times \nabla_p$) is more difficult when the E vs k surfaces are the warped spheres discussed in Chapter 5. The relaxation time is taken as a scalar, because we are now working at $\mathbf{k} = 0$, and τ has the symmetry of the cubic lattice. We shall shortly see that the difficulty is the fact that $\mathbf{G}$ as defined in Eq. (7-38) is an explicit function of the p_i and the step immediately following Problem 7-10 is not possible. The procedure which will be followed in this case is to obtain a from Eq. (7-66) by an *iterative* procedure. If B is very small, an approximate solution is

$$a_0 = e\tau\boldsymbol{\mathcal{E}}\cdot\nabla_p E$$

If we substitute this into the second term of Eq. (7-66), the third term gives, as the next approximation to the value of a,

$$\begin{aligned} a_1 &= e\tau\boldsymbol{\mathcal{E}}\cdot\nabla_p E - e\tau\mathbf{B}\cdot\boldsymbol{\Omega}a_0 \\ &= e\tau\boldsymbol{\mathcal{E}}\cdot\nabla_p E - e^2\tau\mathbf{B}\cdot\boldsymbol{\Omega}(\tau\boldsymbol{\mathcal{E}}\cdot\nabla_p E) \end{aligned}$$

For the next approximation, we substitute a_1 in a similar fashion to get

$$\begin{aligned} a_2 = {} & e\tau\boldsymbol{\mathcal{E}}\cdot\nabla_p E - e^2\tau\mathbf{B}\cdot\boldsymbol{\Omega}(\tau\boldsymbol{\mathcal{E}}\cdot\nabla_p E) \\ & + e^3\tau\mathbf{B}\cdot\boldsymbol{\Omega}[\tau\mathbf{B}\cdot\boldsymbol{\Omega}(\tau\boldsymbol{\mathcal{E}}\cdot\Delta_p E)] \end{aligned} \tag{7-139}$$

Note that τ must remain subject to the various operations, since $\tau = \tau(E)$, and that the iteration is now accurate to terms in B^2, which is a valid approximation for small fields. Hence, we denote a_2 as a. Substituting Eq. (7-37) into Eq. (7-43), dropping the term in f_0 and using $\mathbf{v} = \nabla_p E$, gives

$$\mathbf{J} = (2e/h^3)\int \nabla_p E \; a \; \partial f_0/\partial E \; dV_p \tag{7-140}$$

where a is given by Eq. (7-139).

PROBLEM 7-29 Substitute Eq. (7-139) into Eq. (7-140) to prove that

$$J_i = \sum_j \sigma_{ij0}\mathcal{E}_j + \sum_{j,k} \sigma_{ijk}\mathcal{E}_j B_k + \sum_{j,k,l} \sigma_{ijkl}\mathcal{E}_j B_k B_l + \cdots \tag{7-141}$$

where
$$\sigma_{ij0} = (2e^2/h^3) \int \tau\nabla_{pi}E\nabla_{pj}E(\partial f_0/\partial E)dV_p = \sigma_{ij})_{B=0} \tag{7-142}$$

$$\sigma_{ijk} = -(2e^3/h^3) \int \tau\nabla_{pi}\, E\nabla_{pj}\, E\nabla_{pj}(\tau\nabla_{pk}E)(\partial f_0/\partial E)\, dV_p \tag{7-143}$$

$$\sigma_{ijkl} = (2e^4/h^3) \int \tau\nabla_{pi}\, E\nabla_{pj}\, E\nabla_{pj}[\tau\nabla_{pk}\, E\nabla_{pk}(\tau\nabla_{pl}E)](\partial f_0/\partial E)\, dV_p \tag{7-144}$$

and all indices run from 1 to 3. [Hint: Use the formula from Appendix A2: $(\mathbf{A} \times \mathbf{B})_i = \Sigma\, \epsilon_{ijk}\, A_j B_k$.]

It was shown by Jones (1956)[7] that Eq. (7-144) can be simplified through an integration by parts. Let φ and ψ be two arbitrary functions of the momentum p. Then

$$\mathbf{\Omega}(\varphi\psi) = \varphi(\mathbf{\Omega}\psi) + \psi(\mathbf{\Omega}\varphi) \tag{7-145}$$

PROBLEM 7-30

(a) Prove Eq. (7-145).

(b) Use Fig. 2-4 to prepare a reasonably accurate sketch of $\partial f_0/\partial E$.

(c) Multiply Eq. (7-145) by $(\partial f_0/\partial E)\, dV_p$ and integrate from 0 to ∞. Making use of the properties of $\partial f_0/\partial E$ and of the fact that ∇_p and $\int (\quad) dV_p$ are inverse operations, show that the left-hand side of the resulting expression vanishes, and hence

$$\int_0^\infty \varphi(\mathbf{\Omega}\psi)\partial f_0/\partial E\, dV_p = -\int_0^\infty \psi(\mathbf{\Omega}\varphi)\partial f_0/\partial E\, dV_p \tag{7-146}$$

(d) Apply Eq. (7-146) to Eq. (7-144) and show that

$$\sigma_{ijkl} = (2e^4/h^3) \int \tau\nabla_{pi}\, E\Omega_j[\tau\Omega_k(\tau\nabla_{pl}E)](\partial f_0/\partial E)dV_p \tag{7-147}$$

(e) Show from Eq. (7-142) that

$$\sigma_{ij0} = \sigma_{ji0} \tag{7-148}$$

(f) Show from Eqs. (7-143) and (7-146) that

$$\sigma_{ijk} = -\sigma_{kji}, \qquad \sigma_{iji} = 0 \tag{7-149}$$

(g) Show that Eqs. (7-144) and (7-146) lead to

$$\sigma_{ijkl} = \sigma_{lkji} \tag{7-150}$$

(h) Simplify Eq. (7-77) by letting m and τ be scalars. Show that

$$\Delta = 1 + \omega^2\tau^2$$

and find Δ^{-1} correctly to terms in B^2. Using this result and Eqs. (7-147) through (7-149), show that Eqs. (7-75) and (7-77) combine to give a relation which is of the same form as Eq. (7-141) except for a few terms. What are these terms? Show that the quantities in Eq. (7-148) are related to the conductivity.

Wannier (1959)[8] has shown how the procedure used above can give a simple insight into the relation between the conductivity tensor $\boldsymbol{\sigma}_0$ and the reciprocal effective mass tensor $\boldsymbol{\mu}$. Integrating Eq. (7-142) by parts and assuming that τ is energy-independent, we obtain

$$\sigma_{ij0} = -(2e^2\tau/h^3) \int f_0\, \nabla_{pi}\, \nabla_{pj}\, E\, dV_p$$

It was shown in connection with Problem 7-11 that the quantities $\nabla_{pi}\, \nabla_{pj}\, E$ are the components of $\boldsymbol{\mu}$, so

$$\boldsymbol{\sigma}_0 = -2e^2\tau < \boldsymbol{\mu} > \tag{7-151}$$

where $< >$ denotes the average over the equilibrium distribution and the quantity h^3, a unit of volume in phase-space, is incorporated into the average. This expression should be compared with the scalar form of Eq. (7-8).

In order to use Eqs. (7-142), (7-143), and (7-144) to compute the galvanomagnetic coefficients for P-type silicon and germanium, it is necessary to evaluate the derivatives $\nabla_{pi}\, E$. Lax and Mavroides (1955) have shown how this can be done by using a series approximation. Write Eq. (5-51) as

$$E = -(\hbar^2/2m)(Ak^2 \pm \sqrt{B^2k^4 + C^2\, S}) \tag{7-152}$$

where
$$S = k_1^2k_2^2 + k_2^2k_3^2 + k_3^2k_1^2$$

Herman (1955)[10] states that spheres having a curvature equal to the average curvature of the two warped surfaces have the equation

$$E = -(\hbar^2k^2/2m)(A + \sqrt{B^2 + C^2/6}) \tag{7-153}$$

PROBLEM 7-31 By considering the warped surfaces to be slightly distorted spheres, show that the two sets of surfaces [Eqs. (7-152) and (7-153)] intersect along the cube diagonals.

Equation (7-152) can be written in terms of Eq. (7-153), as follows:

$$E = -(\alpha\hbar^2k^2/2m)[1 - (\gamma S/k^4) - (\gamma/6)] \qquad (7\text{-}154)$$

where

$$B' = (B^2 - C^2/6)^{1/2}$$

$$\gamma = \mp C^2/2B'(A \pm B') = \mp C^2/2B'\alpha$$

$$\alpha = A \pm B'$$

and the upper signs correspond to the positive sign in Eq. (7-152). The quantity γ is called the *anisotropy parameter*, since it indicates the amount of warping.

PROBLEM 7-32 Show that the first neglected term in Eq. (7-154) is $-\delta^2$, where

$$\delta = C^2(S/k^4 - \tfrac{1}{6})/2B'^2$$

and show that its maximum occurs along the cube diagonals. Estimate the amount of error for silicon and germanium using the values of A, B, and C given in connection with Eq. (5-51).

We now convert to spherical coordinates by

$$\left.\begin{aligned} k_1 &= k \sin\theta \cos\varphi \\ k_2 &= k \sin\theta \sin\varphi \\ k_3 &= k \cos\theta \end{aligned}\right\} \qquad (7\text{-}155)$$

PROBLEM 7-33 Show that Eq. (7-127) becomes

$$E = -(\alpha\hbar^2k^2/2m)(1 + \gamma q) \qquad (7\text{-}156)$$

where

$$q = [\sin^4\theta\,(\cos^4\varphi + \sin^4\varphi) + \cos^4\theta - \tfrac{2}{3}]/2$$

The components of the gradient are found by differentiating Eq. (7-154) and then using Eq. (7-155), giving three relations of the form

$$\nabla_{pi}E = \partial E/\partial p_i = (-\alpha/m)p_i(1 + \gamma q_i) \qquad (7\text{-}157)$$

where q_i is a dimensionless polynomial in θ and φ. Similarly,

$$\left.\begin{aligned} \partial^2E/\partial p_i^2 &= (-\alpha/m)(1 + \gamma q_{ii}) \\ \partial^2E/\partial p_i\partial p_j &= 2(\alpha/m)p_ip_j\gamma q_{ij} \\ \partial^3E/\partial p_i\partial p_i\partial p_j &= 2(\alpha/m)p_j\gamma q_{iij}/p^2 \\ \partial^3E/\partial p_i\partial p_j\partial p_k &= 4(\alpha/m)\gamma p_ip_jp_kq_{ijk}/p^4 \end{aligned}\right\} \qquad (7\text{-}158)$$

We now introduce

$$\tau = lE^\lambda \qquad (7\text{-}117)$$

and a variation of Eq. (7-130):

$$\Gamma(y + \tfrac{1}{2})/2a^{y+1/2} = \int_0^\infty x^{2y}e^{-ax^2}\,dx \qquad (7\text{-}159)$$

PROBLEM 7-34 Establish Eq. (7-159) from Eq. (7-130).

We will write Eq. (7-118) as

$$f_0 = e^{(E_V+E-E_F)/kT} \qquad (7\text{-}160)$$

where we are using the top of the valence band as the zero of energy and have adjusted the sign in the exponent so we can integrate from 0 to ∞. A typical element of the form σ_{ii0} then becomes

$$\sigma_{110} = (2e^2/h^3)l \int [(-\alpha/m)p\cos\varphi\sin\theta(1 + \gamma q_1)]^2[(jp^2kT)^\lambda e^{jp^2}]$$
$$[e^{(E_V-E_F)/kT}p^2\,dp\,\sin\theta\,d\theta\,d\varphi/kT] \qquad (7\text{-}161)$$

where

$$j(\theta, \varphi) = (-\alpha/2mkT)(1 + \gamma q)$$

and

$$E = jp^2kT$$

PROBLEM 7-35 Verify Eq. (7-161).

Integrating over p leaves

$$\sigma_{110} = \frac{(-1)^{\lambda+1}(e^2/h^3)e^{(E_V-E_F)/kT}(m/\alpha)^{1/2}(kT)^{\lambda+1.5}2^{5/2}\Gamma(\lambda + 2.5)}{\displaystyle\int (1 + \gamma q_1)^2(1 + \gamma q)^{-5/2}\sin^3\theta\cos^2\varphi\,d\theta\,d\varphi} \qquad (7\text{-}162)$$

PROBLEM 7-36 Expand the first two quantities under the integral sign of Eq. (7-162) and, using term-by-term integration, show that

$$\int_0^\pi\int_0^{2\pi} (1 + \gamma q_1)^2(1 + \gamma q)^{-5/2}\sin^3\theta\cos^2\varphi\,d\theta\,d\varphi$$
$$= \frac{4\pi}{3}\left(1 + \frac{\gamma}{60} + \cdots\right) \qquad (7\text{-}163)$$

Using Eq. (7-163) in Eq. (7-162) and going through the same procedure for the other σ_{ij0} gives

$$\sigma_{ij0} = \frac{(-1)^{\lambda+1}e^2 l\, 2^{9/2} e^{(E_V-E_F)/kT} m^{1/2} (kT)^{\lambda+1.5} \pi \Gamma(\lambda + 2.5)\delta_{ij}}{3h^3\alpha^{1/2}} S_1(\gamma) \qquad (7\text{-}164)$$

where $$S_1(\gamma) = \alpha^{-1/2}(1 + 0.016\gamma + \cdots)$$

Similar calculations for the other coefficients give

$$\sigma_{ijk} = \frac{(-1)^{2\lambda+1}e^3 l^2 2^{9/2} e^{(E_V-E_F)/kT} \alpha^{1/2} (kT)^{2\lambda+1.5} \Gamma(2\lambda + 2.5)\epsilon_{ijk}}{3h^3 m^{1/2}} S_2(\gamma) \qquad (7\text{-}165)$$

where $$S_2(\gamma) = \alpha^{1/2}(1 - 0.016\gamma + \cdots)$$

and
$$\left.\begin{aligned} \sigma_{iiii} &= -\frac{1024}{3465} A\gamma^2 S_3(\gamma) \\ \sigma_{iijj} &= -\frac{64}{3} A\, S_4(\gamma) \\ \sigma_{ijij} &= \frac{64}{3} A\, S_5(\gamma) \\ \sigma_{ijji} &= \frac{512}{3465} A\gamma^2 S_6(\gamma) \end{aligned}\right\} \qquad (7\text{-}166)$$

where $$A = \frac{e^4 e^{(E_V-E_F)/kT} l^3 (kT)^{3\lambda+1.5} \Gamma(3\lambda + 2.5)\alpha^{3/2}}{h^3(2m)^{3/2}}$$

The series $S_3(\gamma)$ through $S_6(\gamma)$ are all different and are given by Lax and Mavroides (1955-56).[11]

The warped energy surfaces require a modification in Eq. (3-19) for the density of holes p in the valence band, as discussed at the beginning of Section 7-4. We can find p from

$$p = (2/h^3) \int f_0\, dV_p \qquad (7\text{-}167)$$

PROBLEM 7-37 Establish Eq. (7-167).

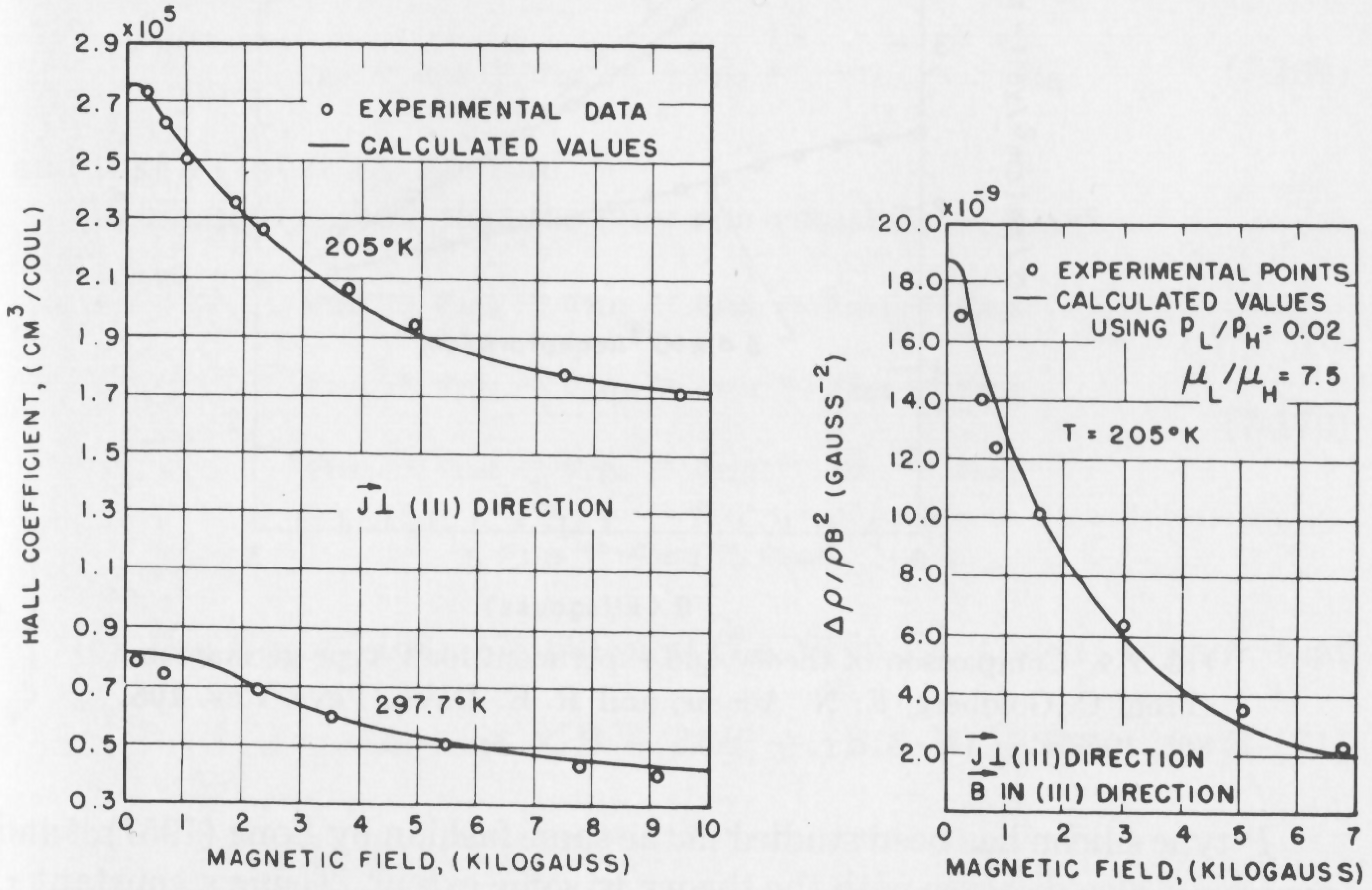
By the same reasoning as above,

$$p = 2(2\pi mkT/h^3)^{3/2} e^{(E_V-E_F)/kT} \alpha^{-3/2} S_0(\gamma)$$

where $$S_0(\gamma) = 1 + 0.05\gamma + \cdots$$

and we note that when γ is small, this reduces to the expression for the spherical case.

The comparison of theory and experiment for P-type crystals has not been as successful as that for N-type. Part of the difficulty may be due to the fact that the constants A, B, and C determining the shape of the valence band warped spheres are determined by cyclotron resonance experiments performed at 4°K, but it is not possible to determine these directly at other temperatures. However, Willardson, Harmon, and Beer (1954)[12] were able to obtain the agreement shown in Figs. 7-8a and 7-8b on the assumption that the energy surfaces are spherical but that there are two kinds of holes. For the magnetoresistivity curves, Fig. 7-8b, p_L/p_H denotes the ratio of light-hole to heavy-hole density, and the values shown on the graph were chosen to make the theoretical curve fit the experimental data.



(a) (b)

FIG. 7-8. Galvanomagnetic coefficients of P-type germanium. (From R. K. Willardson, T. C. Harmon, and A. C. Beer, Phys. Rev. **96,** 1512, 1954, with permission.)

Goldberg, Adams, and Davis (1957)[13] have also measured the Hall coefficient and magnetoresistivity in P-type germanium, and they did not obtain very good agreement between theory and experiment. Figure 7-9 shows the ratio of the conductivity σ in the magnetic field to pe for samples of two different purities, and the theoretical curve as computed from the formulas derived in this section. Their explanation for the discrepancy is that the surfaces at 77°K are considerably different from those at 4°K or that the scattering does not obey a simple relation of the form $\tau = lE^{\lambda}$.

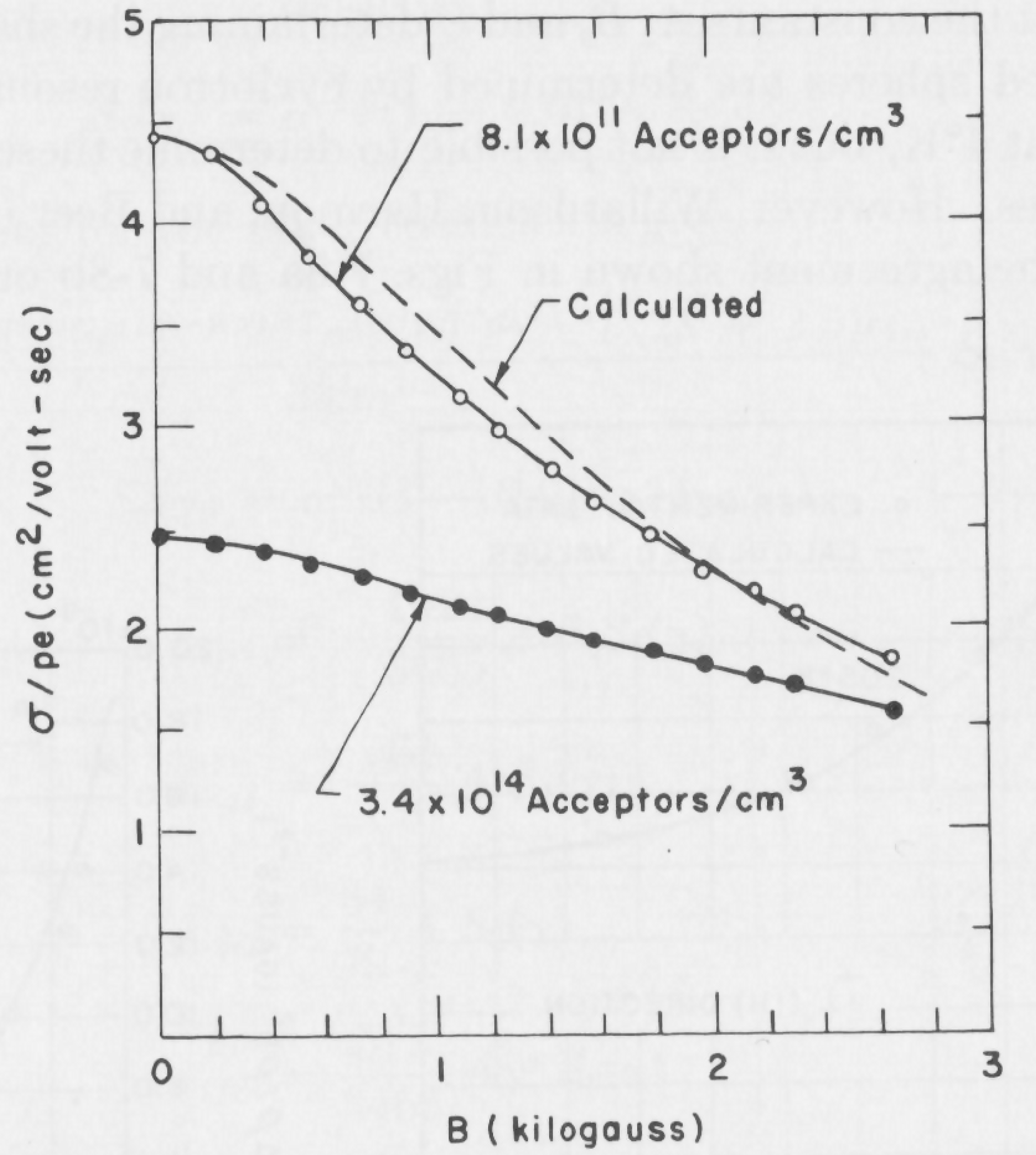


FIG. 7-9. Comparison of theory and experiment for P-type germanium. (From C. Goldberg, E. N. Adams, and R. E. Davis, Phys. Rev. **105**, 865, 1957.)

P-type silicon has been studied in the same fashion by Long (1957),[14] and his results also disagree with the theory to some extent. Using a constant τ, Lax and Mavroides[9] compute the ratio of μ_H/μ as 1.76 for germanium and 1.24 for silicon, and for $\lambda = \frac{1}{2}$ the respective values are 2.07 and 1.45; whereas Long[14] measures this ratio as 0.75 at room temperature, increasing to 1.0 at 77°K.

We shall conclude this section by introducing an alternative form of Eq. (7-141), applicable to cubic crystals, which is due to Seitz (1950).[16] The energy surfaces in silicon and germanium, ellipsoids for the conduction band and warped spheres for the valence band, are invariant under the symmetry operations of the cube. This concept can be applied to the integrals of Eqs. (7-142), (7-143), and (7-147) to obtain further information about the coefficients in Eq. (7-141). If we choose the coordinate system *OXYZ* parallel to the crystal axes and assume that τ has the same symmetry as the energy surfaces, then the integrations, because of this symmetry, involve even or odd functions only. In particular, the gradient components

$\nabla_{pi}E$ are odd functions, and the components Ω_i of Ω are even. Using the fact that the integral of an odd function over the range $\infty < p_i < -\infty$ will vanish, and taking into consideration the cubic symmetry, we see that

$$\left.\begin{aligned} \sigma_{rs0} &= 0 \qquad \text{for } r \neq s \\ \sigma_{110} &= \sigma_{220} = \sigma_{330} = \sigma_0 \end{aligned}\right\} \tag{7-168}$$

PROBLEM 7-38

(a) Apply the reasoning above to show that, for cubic crystals,

$$\sigma_{123} = \sigma_{231} = \sigma_{312} = -\sigma_{132} = -\sigma_{213} = -\sigma_{321} \tag{7-169}$$

and that all other σ_{ijk} are zero.

(b) Similarly, show that the only non-vanishing σ_{ijkl} are

$$\left.\begin{aligned} \sigma_{1221} &= \sigma_{2112} = \sigma_{3113} = \sigma_{1331} = \sigma_{2332} = \sigma_{3223} \\ \sigma_{1212} &= \sigma_{2121} = \sigma_{1313} = \sigma_{3131} = \sigma_{2323} = \sigma_{3232} \\ \sigma_{1122} &= \sigma_{1133} = \sigma_{2233} = \sigma_{2211} = \sigma_{3311} = \sigma_{3322} \\ \sigma_{1111} &= \sigma_{2222} = \sigma_{3333} \end{aligned}\right\} \tag{7-170}$$

(c) Use Eqs. (7-168), (7-169), and (7-170) in Eq. (7-141) to show that

$$\mathbf{J} = \sigma_0\boldsymbol{\mathcal{E}} + \alpha\boldsymbol{\mathcal{E}} \times \mathbf{B} + \beta\boldsymbol{\mathcal{E}}B^2 + \gamma\mathbf{B}(\boldsymbol{\mathcal{E}}\cdot\mathbf{B}) + \delta\boldsymbol{\theta}\boldsymbol{\mathcal{E}} \tag{7-171}$$

where $\alpha = \sigma_{123}, \quad \beta = \sigma_{1221}$, etc.

and $\boldsymbol{\theta} = \begin{bmatrix} B_1^2 & 0 & 0 \\ 0 & B_2^2 & 0 \\ 0 & 0 & B_3^2 \end{bmatrix} = B_iB_j\delta_{ij}$

(d) Let $B_1 = B_2 = 0$, $\mathcal{E}_3 = 0$, and $J_2 = J_3 = 0$. Show that

$$J_1 = \sigma_0\mathcal{E}_1 + \sigma_{123}\mathcal{E}_2B_3 + \sigma_{1331}\mathcal{E}_1B_3^2 \tag{7-172}$$

$$0 = \sigma_0\mathcal{E}_2 + \sigma_{123}\mathcal{E}_1B_3 + \sigma_{2332}\mathcal{E}_2B_3^2 \tag{7-173}$$

Solve Eq. (7-173) for $\mathcal{E}_2$, and substitute this result plus Eq. (7-172) into Eq. (7-12). Dropping terms in B^2, show that

$$R_H = -\sigma_{123}/\sigma_0^2 \tag{7-174}$$

(e) Let $B_1 = B_2 = J_1 = J_2 = 0$. Show that

$$J_3 = \mathcal{E}_3[\sigma_0 + B_3^2(\beta + \gamma + \delta)] \qquad (7\text{-}175)$$

and

$$M_l = B_3^2(\beta + \gamma + \delta) \qquad (7\text{-}176)$$

(f) Show that Eq. (7-171) can be solved for $\boldsymbol{\mathcal{E}}$ to give

$$\boldsymbol{\mathcal{E}} = \rho_0\mathbf{J} + R_H\mathbf{J} \times \mathbf{B} + \rho_0[bB^2\mathbf{J} + c\mathbf{B}(\mathbf{J}\cdot\mathbf{B}) + d\boldsymbol{\theta}\,\mathbf{J}] \qquad (7\text{-}177)$$

7-9 The Electrical Properties of Tellurium

Some of the most interesting (and baffling) properties of tellurium are (1) the double reversal of the Hall coefficient as a function of temperature and (2) the fact that it is always P-type. The Hall coefficient normally reverses in P-type semiconductors; the reason for this can be seen from the formula

$$R = 3\pi(p - nb^2)/8e(p + nb)^2 \qquad (7\text{-}20)$$

Below the intrinsic temperature, conduction is predominantly by holes, and Eq. (7-20) reduces to $R = 3\pi/8pe$ when $p \gg n$. As the material is warmed up, electrons are excited into the conduction band; because their mobility is generally higher than that of holes, they make the material effectively N-type even when $n < p$. This can be seen by equating R to zero, so that

$$p = nb$$

is the condition for the reversal of the Hall coefficient. Hence, when $n > p/b$, R is negative; this reversal in sign occurs in pure tellurium at about −20°C. Figure 7-10 of Nussbaum and Hager (1961)[15] shows this transition from P-type to N-type, as well as a reversal back to N-type at about 260°C (the curve taken with pressure applied to the sample should be disregarded for the moment). This anomalous second reversal was discovered by Wold (1916).[17] The cause is not positively understood, although a possible explanation will be given shortly.

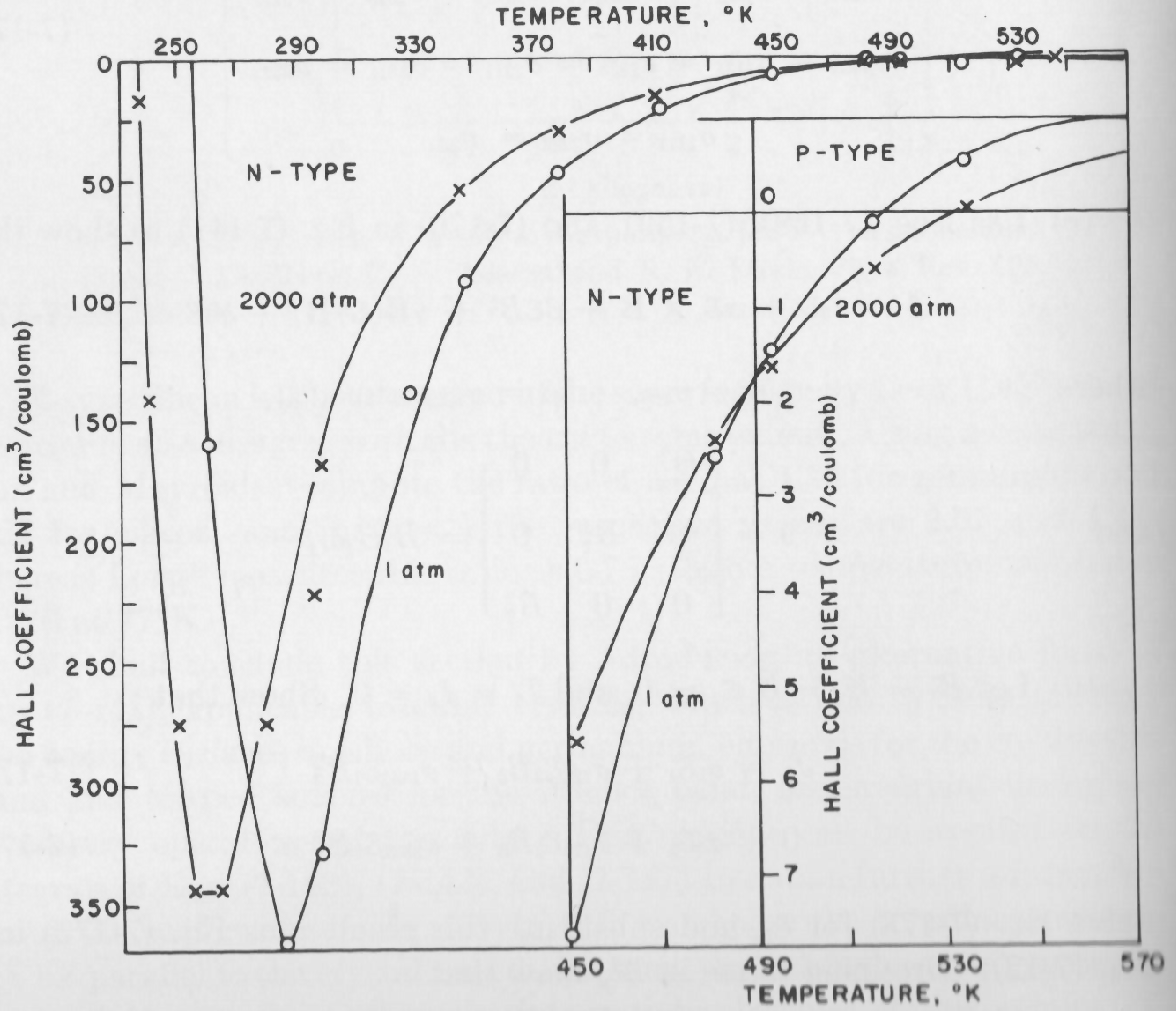


FIG. 7-10. Hall coefficient as a function of temperature and pressure in tellurium. (From A. Nussbaum and R. J. Hager, Phys. Rev. **123**, 1958, 1961.)

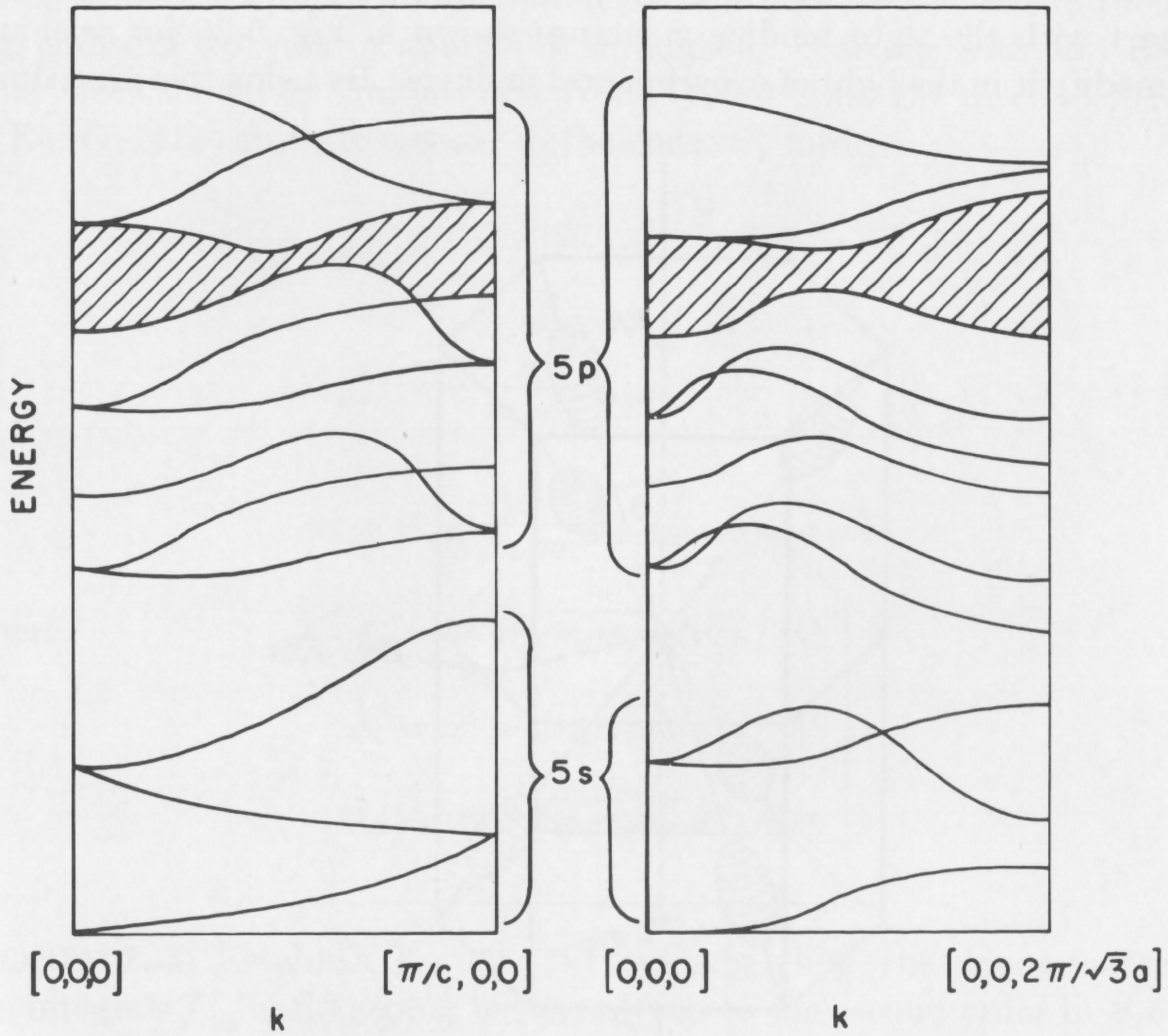


FIG. 7-11. Energy band structure of tellurium. (From Nussbaum and Hager, *op. cit.*)

A thorough discussion of the physical, electrical, and optical properties of tellurium has been given in a review article by Blakemore et al. (1961,[18] so we shall concentrate here on the galvanomagnetic coefficients and their relation to the band structure. Precise calculations of the kind leading to Fig. 5-37 have not been made for tellurium, and the best that can be done is to start with the tight-binding structure shown in Fig. 5-36 for selenium and modify it in the light of experimental findings. By using this procedure,

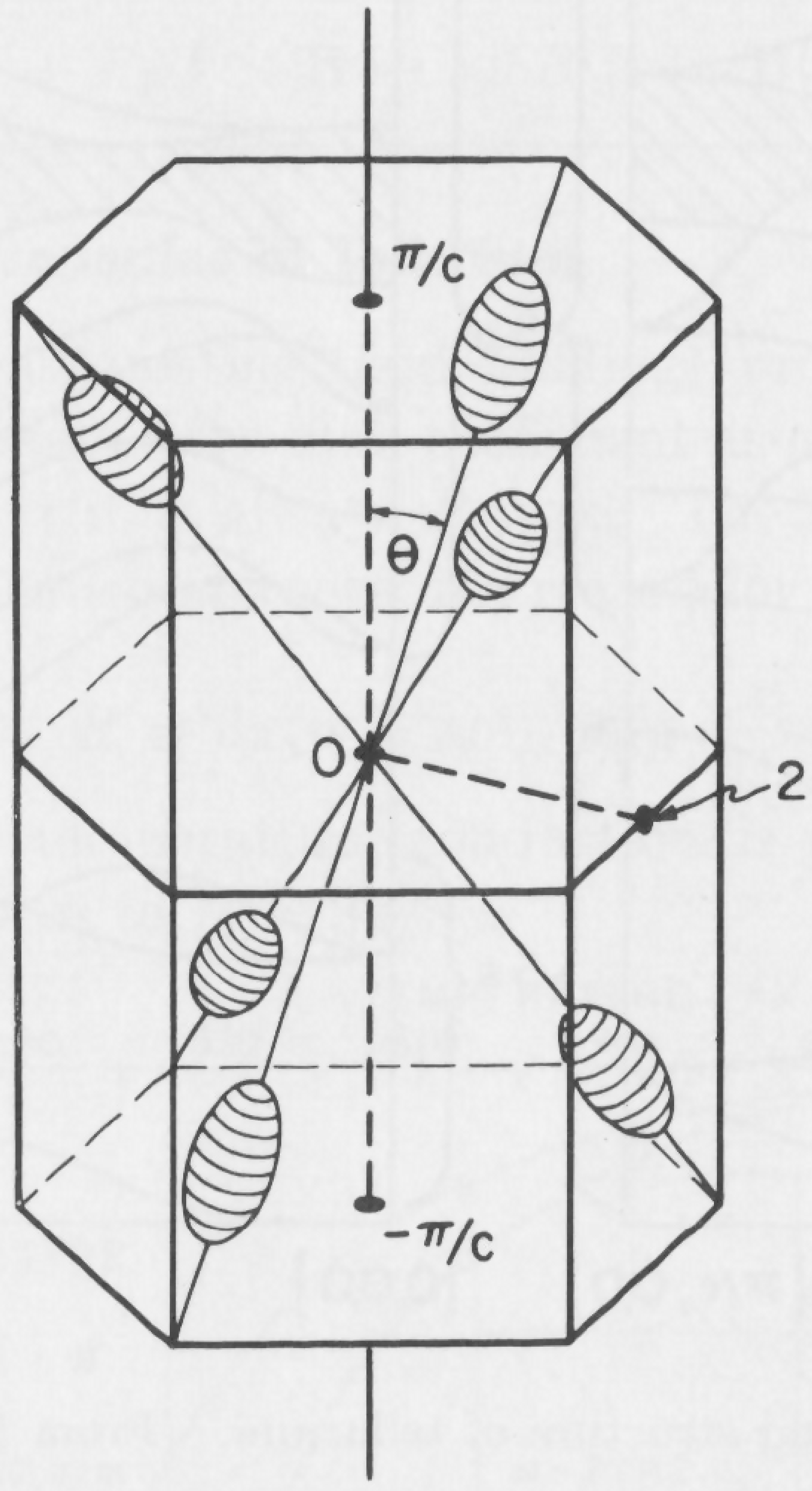


FIG. 7-12. The arrangement of the energy ellipsoids in tellurium. (From Nussbaum and Hager, *op. cit.*)

Nussbaum and Hager (1961)[15] arrived at the proposed band picture of Fig. 7-11; the arguments for the structure shown will be deferred until we have discussed the calculation of the galvanomagnetic tensor. The E vs k curves show a minimum in the conduction band and a maximum in the valence band in both Figs. 7-11a and 7-11b, giving ellipsoids at an interior point of the Brillouin zone. We must realize that these curves are actually two-dimensional representations of the behavior of E as we travel through the Brillouin zone; the symmetry requires that if the constant energy surfaces are ellipsoids, they must occur in six equivalent places, as shown in Fig. 7-12.

PROBLEM 7-39 Show that the Brillouin zone for tellurium has the shape indicated in Fig. 7-12, and verify that the positions of the ellipsoids satisfy the symmetry properties. Do these ellipsoids have axial symmetry?

Let us consider now the conductivity tensor associated with one of these six equivalent ellipsoids and denote it by $\boldsymbol{\sigma}^j$, where the superscript denotes the ellipsoid (or valley) number 1 through 6. Starting with Eq. (7-77), it has been shown by Drabble and Wolfe (1956)[19] that the three coefficients in Eq. (7-141) can be expressed in the compact form

$$\left.\begin{aligned} \sigma^j_{pqo} &= \frac{A_1}{m_p} \\ \sigma^j_{pqr} &= -\frac{A_2}{m_p m_q}\,\epsilon_{pqr} \\ \sigma^j_{p\,qrs} &= \frac{A_3}{m_p m_q m_t}\,[\epsilon_{tps}\epsilon_{qtr} + \epsilon_{tpr}\epsilon_{qts}] \end{aligned}\right\} \tag{7-178}$$

where

$$\left.\begin{aligned} A_1 &= 2^{9/2}e^2\sqrt{m_1 m_2 m_3}\,\pi I_{11}/3h^2 \\ A_2 &= 2^{9/2}e^3\sqrt{m_1 m_2 m_3}\,\pi I_{12}/3h^2 \\ A_3 &= 2^{9/2}e^4\sqrt{m_1 m_2 m_3}\,\pi I_{13}/6h^2 \end{aligned}\right\} \tag{7-179}$$

PROBLEM 7-40 Establish Eq. (7-178) by expanding the denominators of the integrals I^j_{1n} and keeping terms only up to the second order in B_iB_j.

The matrix for transforming the coordinate axes $OXYZ$ of the ellipsoid to the set $O'X'Y'Z'$ of the crystal by Fig. 7-13 and Appendix A4, is

$$\mathsf{A} = \begin{pmatrix} C & 0 & -S \\ 0 & 1 & 0 \\ S & 0 & C \end{pmatrix} \tag{7-180}$$

where $C = \cos\theta \qquad S = \sin\theta$

The corresponding matrices for the other five ellipsoids can be found by multiplying this one by the matrices representing the symmetry properties of the ellipsoidal arrangement.

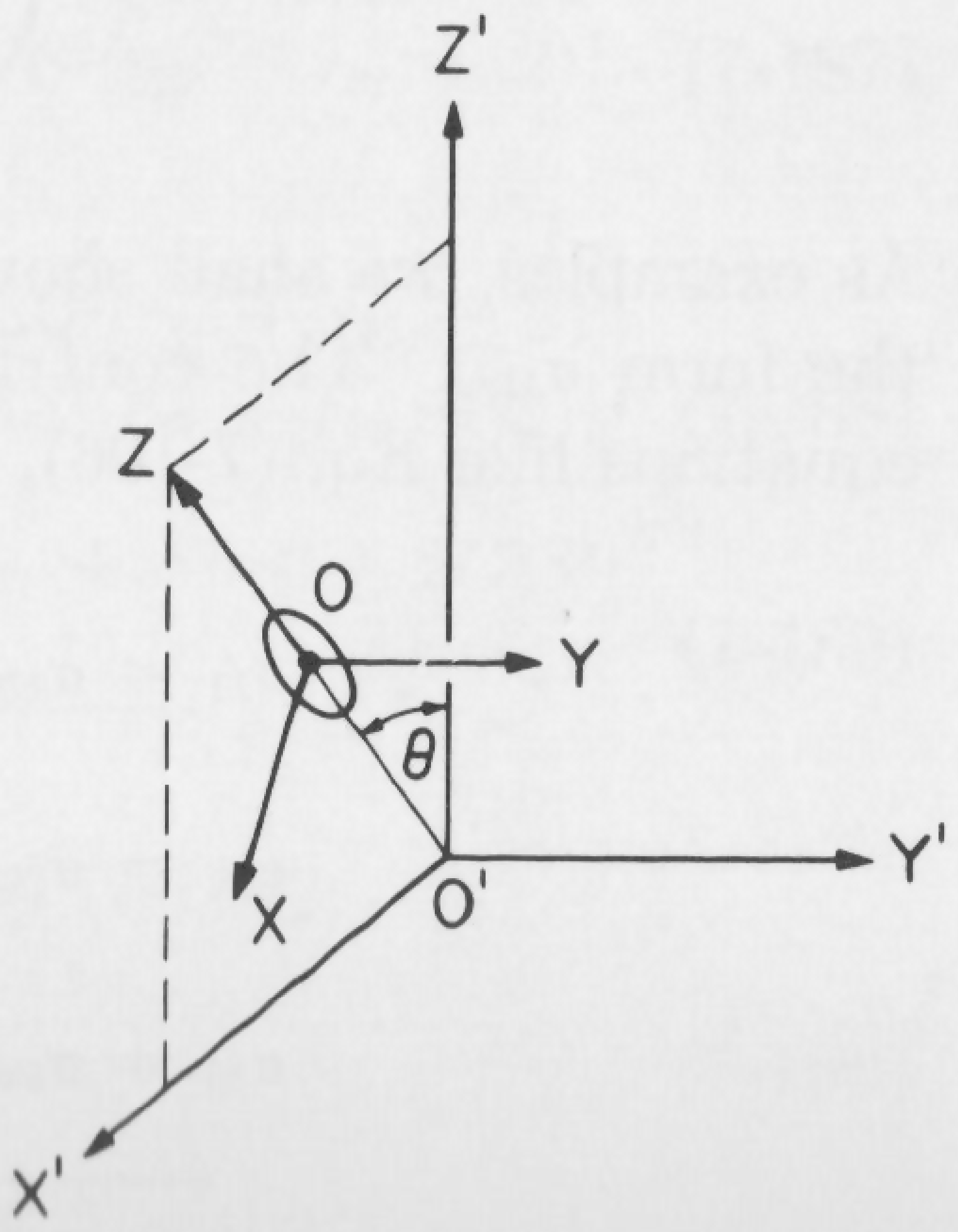


FIG. 7-13. Transformation from ellipsoidal to crystal axes in tellurium.

PROBLEM 7-41 Show that the remaining five matrices are

$$\begin{pmatrix} -C/2 & -\sqrt{3}C/2 & -S \\ \sqrt{3}/2 & -\frac{1}{2} & 0 \\ -S/2 & -\sqrt{3}S/2 & C \end{pmatrix} \tag{7-181}$$

$$\begin{pmatrix} -C/2 & \sqrt{3}C/2 & -S \\ -\sqrt{3}/2 & -\frac{1}{2} & 0 \\ -S/2 & \sqrt{3}S/2 & C \end{pmatrix} \tag{7-182}$$

$$\begin{pmatrix} C & 0 & S \\ 0 & -1 & 0 \\ S & 0 & -C \end{pmatrix} \tag{7-183}$$

$$\begin{pmatrix} -C/2 & -\sqrt{3}C/2 & S \\ -\sqrt{3}/2 & \frac{1}{2} & 0 \\ -S/2 & -\sqrt{3}S/2 & -\mathrm{C} \end{pmatrix} \tag{7-184}$$

$$\begin{pmatrix} -C/2 & \sqrt{3}C/2 & S \\ \sqrt{3}/2 & \frac{1}{2} & 0 \\ -S/2 & \sqrt{3}S/2 & -C \end{pmatrix} \tag{7-185}$$

As examples, we shall show how to calculate magnetoresistivity terms of the form σ_{11rs}. The contributions of each ellipsoid are, again, given by equations like Eq. (7-106), and Eqs. (7-178) and (7-179) show that

$$\sigma_{11} = \sigma_{110} + \sigma_{1122}B_2^2 + \sigma_{1133}B_3^2$$

$$\sigma_{22} = \sigma_{220} + \sigma_{2211}B_1^2 + \sigma_{2233}B_3^2$$

$$\sigma_{33} = \sigma_{330} + \sigma_{3311}B_1^2 + \sigma_{3322}B_2^2$$

$$\sigma_{13} = -\sigma_{132}B_2 + \sigma_{1313}B_1B_3 + \sigma_{1331}B_3B_1$$

$$\sigma_{31} = \sigma_{312}B_2 + \sigma_{3131}B_3B_1 + \sigma_{3113}B_1B_3$$

Then the sum of all the contributions to σ'_{11rs} is

$$\left.\begin{array}{rllllll} \sigma'_{11rs} = C^2\sigma_{1122}(& \overset{①}{B}{}_2^2 & +\overset{②}{B}{}_2^2/4 & +\overset{③}{B}{}_2^2/4 & +\overset{④}{B}{}_2^2 & +\overset{⑤}{B}{}_2^2/4 & +\overset{⑥}{B}{}_2^2/4) \\ +\,C^2\sigma_{1133}(& B_3^2 & +B_3^2/4 & +B_3^2/4 & +B_3^2 & +B_3^2/4 & +B_3^2/4) \\ +\,(\frac{3}{4})\sigma_{2211}(& 0 & +B_1^2 & +B_1^2 & +0 & +B_1^2 & +B_1^2) \\ +\,(\frac{3}{4})\sigma_{2233}(& 0 & +B_3^2 & +B_3^2 & +0 & +B_3^2 & +B_3^2) \\ +\,S^2\sigma_{3311}(& B_1^2 & +B_1^2/4 & +B_1^2/4 & +B_1^2 & +B_1^2/4 & +B_1^2/4) \\ +\,S^2\sigma_{3322}(& B_2^2 & +B_2^2/4 & +B_2^2/4 & +B_2^2 & +B_2^2/4 & +B_2^2/4) \\ +4\,CS\,\sigma_{1313}(& B_1B_3 & +B_1B_3/4 & +B_1B_3/4 & +B_1B_3 & +B_1B_3/4 & +B_1B_3/4) \\ +\,(\sqrt{3}C)\,\sigma_{1212}(& 0 & -B_1B_2 & +B_1B_2 & +0 & -B_1B_2 & +B_1B_2) \\ +\,(\sqrt{3}S)\,\sigma_{2323}(& 0 & -B_2B_3 & +B_2B_3 & +0 & -B_2B_3 & +B_2B_3) \end{array}\right\} \tag{7-186}$$

where the little numbers in circles indicate that the B_i also must be transformed and the transformation is different for each ellipsoid. That is,

$$\overset{①}{B}_1 = \sum A_{1i}\overset{①}{B}{}'_i, \text{ etc.}$$

or

$$\overset{①}{B}_1 = (CB'_1 + 0B'_2 - SB'_3)$$

Hence, the coefficient of $B_1'^2$ in B_1^2 is C^2 for ellipsoid ①. Thus, we obtain (dropping the primes)

$$\begin{aligned}\sigma_{1111} = (-3A_3/2m_1m_2m_3)[3C^2S^2(r_{31} + r_{13} - 2) + C^2(r_{21} + r_{12} - 2) \\ + S^2(r_{23} + r_{32} - 2)]\end{aligned} \tag{7-187}$$

where $r_{ij} = m_i/m_j$. Similarly,

$$\begin{aligned}\sigma_{1122} = (-3A_3/2m_1m_2m_3)[C^2S^2(r_{31} + r_{13} - 2) + 3C^2(r_{21} + r_{12} - 2) \\ + 3S^2(r_{23} + r_{32} - 2)]\end{aligned} \tag{7-188}$$

$$\sigma_{1133} = (-6A_3/m\ m\ m)[C^4r_{31} + S^4r_{13} + C^2r_{32} + S^2r_{12} + 2C^2S^2] \tag{7-189}$$

PROBLEM 7-42

(a) Verify Eqs. (7-186), (7-187), (7-188), and (7-189).

(b) Show that

$$\sigma_{1123} = 0 \tag{7-190}$$

Roth (1959)[20,21] has shown that a crystal belonging to the point group D_3 normally has 12 non-vanishing galvanomagnetic coefficients, of which two are resistivities, two are Hall coefficients, and eight are magnetoresistivity components. The tensor associated with σ_{ij0} is

$$\boldsymbol{\sigma}_0 = \begin{bmatrix} \sigma_{110} & \sigma_{120} & \sigma_{130} \\ \sigma_{210} & \sigma_{220} & \sigma_{230} \\ \sigma_{310} & \sigma_{320} & \sigma_{330} \end{bmatrix}$$

By the Onsager relations

$$\sigma_{ij0}(B) = \sigma_{ij0}(-B)$$

so this simplifies to

$$\boldsymbol{\sigma}_0 = \begin{bmatrix} \sigma_{110} & \sigma_{120} & \sigma_{130} \\ \sigma_{120} & \sigma_{220} & \sigma_{230} \\ \sigma_{130} & \sigma_{230} & \sigma_{330} \end{bmatrix} \tag{7-191}$$

Since the six ellipsoids are equivalent, any transformation which converts one ellipsoid into another must leave these tensors unchanged. Consider, for example, the effect of a 180-deg rotation about an axis in the XOY-plane of the Brillouin zone. This converts an ellipsoid above the plane into one directly below it, and it may have the matrix

$$\mathsf{A} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{pmatrix}$$

The two ellipsoids are connected by a transformation like Eq. (7-105), and Eq. (7-191) then becomes

$$\boldsymbol{\sigma}_0' = \begin{pmatrix} \sigma_{110}' & -\sigma_{120}' & -\sigma_{130}' \\ -\sigma_{120}' & \sigma_{220}' & \sigma_{230}' \\ -\sigma_{130}' & \sigma_{230}' & \sigma_{330}' \end{pmatrix} \tag{7-192}$$

The only way for this to be the same as Eq. (7-191) is for

$$\sigma_{120}' = \sigma_{130}' = 0$$

PROBLEM 7-43 By going through the same process for the other two-fold axes, show that σ_0 becomes

$$\boldsymbol{\sigma}_0 = \begin{pmatrix} \sigma_{110} & 0 & 0 \\ 0 & \sigma_{110} & 0 \\ 0 & 0 & \sigma_{330} \end{pmatrix}$$

Roth shows in a similar fashion that σ_{pqr} has two non-zero components and σ_{pqrs} has eight. Drabble and Wolfe show that some of these are

$$\sigma_{11} = 3A_1\left(\frac{C^2}{m_1} + \frac{1}{m_2} + \frac{S^2}{m_3}\right)$$

$$\sigma_{33} = 6A_1\left(\frac{S^2}{m_1} + \frac{C^2}{m_3}\right)$$

$$\sigma_{123} = \frac{-6A_2}{m_1m_2m_3}\,(S^2m_1 + C^2m_3)$$

$$\sigma_{231} = \frac{-3A_2}{m_1m_2m_3}\,(C^2m_1 + m_2 + S^2m_3)$$

$$\sigma_{3333} = \frac{-12A_3}{m_1m_2m_3}\,[C^2S^2(r_{31} + r_{13} - 2)]$$

$$\sigma_{1313} = \frac{-6A_3}{m_1m_2m_3}\,[1 + S^2C^2(r_{31} + r_{13} - 2)]$$

and that by inverting the associated tensors, the components of $\boldsymbol{\rho}$ are

$$\begin{aligned} \rho_{11} &= 1/\sigma_{11} & \rho_{1122} &= -\sigma_{1122}/\sigma_{11}^2 - \sigma_{231}^2/\sigma_{11}^2\sigma_{33} \\ \rho_{33} &= 1/\sigma_{33} & \rho_{1133} &= -\sigma_{1133}/\sigma_{11}^2 - \sigma_{123}^2/\sigma_{11}^3 \\ \rho_{123} &= \sigma_{123}/\sigma_{11}^2 & \rho_{3311} &= -\sigma_{3311}/\sigma_{33}^2 - \sigma_{231}^2/\sigma_{11}\sigma_{33}^2 \\ \rho_{231} &= \sigma_{231}/\sigma_{11}\sigma_{33} & \rho_{1123} &= -\sigma_{1123}/\sigma_{11}^2 \\ \rho_{1111} &= -\sigma_{1111}/\sigma_{11}^2 & \rho_{2311} &= -\sigma_{2311}/\sigma_{11}\sigma_{33} \\ \rho_{3333} &= -\sigma_{3333}/\sigma_{33}^2 & \rho_{2323} &= -\sigma_{2323}/\sigma_{11}\sigma_{33} + \tfrac{1}{2}\,\frac{\sigma_{123}\sigma_{231}}{\sigma_{11}^2\sigma_{33}} \end{aligned} \tag{7-193}$$

PROBLEM 7-44

(a) Verify Eq. (7-193).

(b) Gives a simple rule for finding the inverse relations.

It is interesting to note that while the symmetry D_3 indicates eight non-vanishing magnetoresistivities, the further requirement of ellipsoidal energy surfaces reduces this number to six. This has been verified experimentally by Nussbaum and Long (1961).[22] Measurements made at liquid helium, where conduction is due almost entirely to holes, on properly oriented samples of single-crystal tellurium show both ρ_{1123} and ρ_{2311} to be vanishing within the limits of experimental error. Table 7-1 lists measured values of the twelve coefficients at liquid nitrogen and liquid helium temperatures. It is interesting to note that it requires three samples to obtain the twelve elements; these are oriented with the sample axis per-

TABLE 7-1

Coefficient	*Sample Type*	*Data of H. Roth*[19] *77°K, 5000 gauss*	*Crystal 2-8-60*[20] *4.2°K, 1000 gauss*	*Crystal 3-2-60*[20] *4.2°K, 1000 gauss*
ρ_1	A	52 ohm-cm	34 ohm-cm	22 ohm-cm
ρ_3	B	16	16.5	6
R_1	B	2.6×10^4 cm³/coulomb	10.5×10^4 cm³/coulomb	13.5×10^4 cm³/coulomb
R_3	A	7.0×10^4	16×10^4	13×10^4
ρ_{1111}/ρ_1	A	7.2×10^{-10} gauss^{-2}	5.1×10^{-8} gauss^{-2}	6.5×10^{-8} gauss^{-2}
ρ_{1133}/ρ_1	A	8.7×10^{-10}	4.0×10^{-8}	5.2×10^{-8}
ρ_{1122}/ρ_1	A	12×10^{-10}	7.4×10^{-8}	9.5×10^{-8}
ρ_{3333}/ρ_3	B	2.2×10^{-10}	0.8×10^{-8}	3.8×10^{-8}
ρ_{3311}/ρ_3	B	7.2×10^{-10}	3.0×10^{-8}	16.5×10^{-8}
$2\rho_{1313}/(\rho_1 + \rho_3)$	C	-0.35×10^{-10}	0	0
ρ_{1123}/ρ_1	A	-0.4×10^{-10}	0	0
ρ_{2311}/ρ_3	B	0	0	0

pendicular, parallel, and at a 45° angle to the c-axis, respectively, and are denoted by A, B, and C in the table. For simplicity, the abbreviations $\rho_1 = \rho_{11}$, $\rho_3 = \rho_{33}$, $R_1 = \rho_{231}$, and $R_3 = \rho_{123}$ are used.

PROBLEM 7-45 Show that

$$\frac{3\rho_{1111}}{\rho_1^2} - \frac{\rho_{1122}}{\rho_1^2} - \frac{R_1^2}{\rho_1^2\rho_3} = \frac{\rho_{3333}}{\rho_3^2} \tag{7-194}$$

The values of ρ_{3333}/ρ_3^2 as calculated from Eq. (7-194) compare to the measured values as shown in Table 7-2. The agreement is off by a factor of 2 or 5, which shows that it is difficult to draw conclusions based on data from two different samples, even when they are cut from adjacent regions in the same crystal.

TABLE 7-2

Crystal	ρ_{3333}/ρ_3^2 *calculated from* Eq. (7-194)	ρ_{3333}/ρ_3^2 *measured*
2-8-60	0.25×10^8	0.048×10^8
3-2-60	0.34×10^8	0.63×10^8

Returning now to the effect of pressure on the Hall coefficient shown in Fig. 7-10 (this curve is for R_1), we see that at the lower reversal, the application of pressure lowers the temperature at which the reversal occurs. Another anomalous property of tellurium is the fact that when pressure is applied, the chains are squeezed and tend to straighten out, so that the lattice constant c *increases*, whereas the constant a decreases in the normal fashion. The a/c ratio is 0.88 in selenium and 0.75 in tellurium, and pressure, therefore, makes the lattice even more tellurium-like. The band structure of Reitz corresponds to the selenium gap of approximately 2 eV, whereas in tellurium it is only 0.33 eV; comparing the two structures (Figs. 5-36 and 7-11) shows that the thermal gap gets even smaller as the a/c ratio becomes more like tellurium. This makes the Hall coefficient reverse at a lower temperature, since it is easier for the electrons with higher mobility to make the material N-type. The decrease in the energy gap with pressure has been actually measured by Long (1956),[24] who found a 10 per cent change with a pressure of 2000 atm.

The band picture proposed for tellurium can be used to explain the second reversal by postulating a lower-lying hole band with a higher mobility, analogous to the situation in P-type germanium and silicon. However, the energy surfaces here are ellipsoidal and the second band lies far enough below the first one so that it makes no contribution to the conductivity until the temperature reaches the vicinity of 300°K. As the temperature goes higher, the faster holes can cause the crystal to go back to P-type. When pressure is applied in the region of the upper reversal, a new effect comes into play. It is believed that the anisotropic combination of contraction and elongation results in a shifting of the number of states and the density of carriers associated with each ellipsoid, and this affects all the galvanomagnetic coefficients. This process is believed to be the cause of the *piezoresistance* effect — the change in the resistivity of a semiconductor with the application of pressure. In the case of tellurium, the effect of pressure can be accounted for by assuming that the redistribution of carriers among the ellipsoids tends to make less states available for electrons activated from the lower valence band.

Since we are primarily interested in tellurium because of what it teaches us about the band structure and electrical properties of semiconductors, it is worthwhile to mention a suggestion of Callen (1954).[23] By straightening the chains and distorting the lattice angles between atoms on each plane in Fig. 5-25, tellurium can be converted into a tetragonal crystal. This structure is an approximation to the true lattice, but it permits an easy qualitative determination of the band picture. The way this is done is to use the atomic orbitals of Fig. 5-32, with the z-axis taken as the four-fold axis and the other two axes along lattice directions. Let us start with a group of widely spaced tellurium atoms and bring them together to form the tetragonal crystal. The valence electrons in tellurium have the configuration $5s^25p^4$, and the individual atomic s, p, and d levels will spread out, just as the levels do in Fig. 1-23. As the atoms get closer together, the first orbitals to overlap are the d_{z^2} functions of Fig. 5-32, and the corresponding bands are the first to widen. Figure 7-14 illustrates this. In order to show the overlapping in directions parallel and perpendicular to the z-axis on one diagram, we use the *reduced* lattice constant — the ratio of the lattice constant at any separation to its value in the tetragonal crystal — as the variable. The next group of functions to overlap are the $d_{x^2-y^2}$ orbitals, and these form a narrower, higher-lying band. The other d orbitals do not overlap until the atoms are very close together, and consequently do not enter the argument. For the p-bands, the first orbitals to overlap are the p_z functions, which are seen to spread out so much that, at the true lattice constant, they need not be considered. The bottom edges of all the bands are shown as reaching a minimum and then turning up again, a different effect from that of Fig. 1-23. The reason for this comes out of more elaborate calculations than the approximation of Chapter 1. The next p-bands are due to the degenerate p_x and p_y functions.

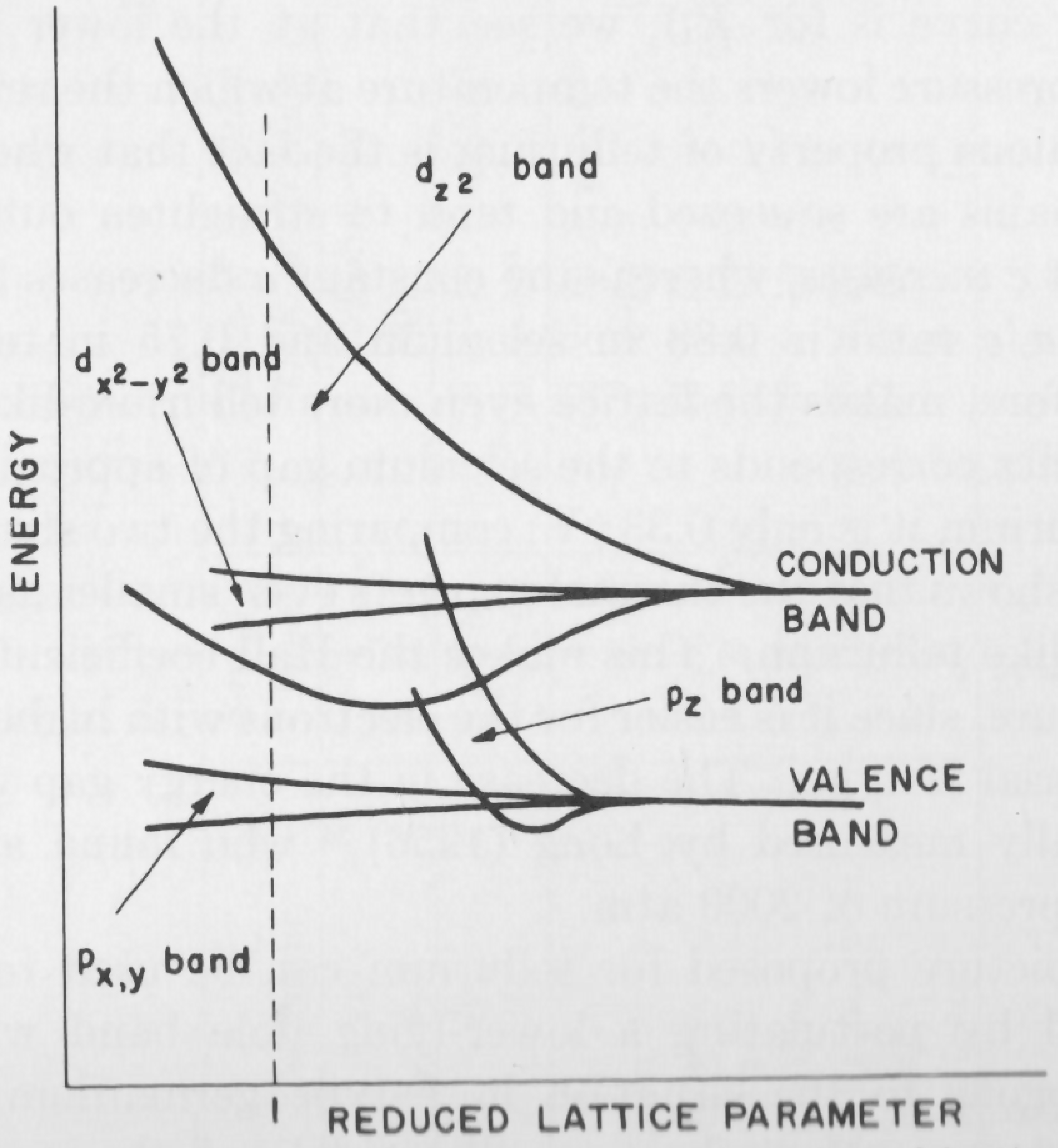


FIG. 7-14. Band structure of tellurium as a function of lattice spacing. (From H. B. Callen, J. Chem. Phys. 22, 518, 1954, with permission.)

PROBLEM 7-46 Sketch the s-bands and explain how they fit into this band picture.

The band structure of Fig. 7-14 shows a double conduction band; the upper reversal can be explained by the presence of the higher-lying, low-mobility electron band. As the temperature is raised, more of the electrons are activated into this band and fewer into the lower one, so the holes can eventually dominate the conductivity. The effect of pressure forces the p_z and d_{z^2} orbitals farther apart (these correspond to c-axis directions in tellurium) and the p_x, p_y, and $d_{x^2-y^2}$ orbitals closer together. On the diagram, this corresponds to a decrease in the gap, agreeing with the results of Long (1956).[24] This picture, therefore, shows what can be done in a qualitative way, but the objection has been raised that the d_{z^2} functions in this simple tetragonal crystal would overlap so much that the material would actually be a metal.

7-10 The Thermomagnetic Effects

As the final topic in this study of transport effects, we shall briefly discuss the phenomena associated with the presence of both an electric and a thermal gradient in a magnetic field. We have seen that a magnetic field acting on the current **J** produces a transverse Hall field. The charge carriers responsible for the Hall voltage can also carry thermal energy and can produce a transverse temperature difference, which is known as the *Ettingshausen effect.* If we replace the electric current **J** with a thermal current $\mathbf{J}_Q$, the magnetic field will still deflect the carriers and will produce transverse voltage and temperature gradients. This results in the *Nernst* and *Righi-Leduc* effects, respectively; all four are summarized in Table 7-3, where the lower two items are known as *thermomagnetic effects.*

TABLE 7-3

	Resultant voltage gradient	*Resultant temperature gradient*
Electric current	Hall	Ettingshausen
Thermal current	Nernst	Righi-Leduc

If the magnetic field is along the z-axis, the Onsager equations, as given by Callen (1952),[25] have the following form:

$$\left.\begin{aligned}
-J_x &= L_{11}\,\nabla_x\bar{\mu} + L_{12}\frac{1}{T}\,\nabla_xT + L_{13}\,\nabla_y\bar{\mu} + L_{14}\frac{1}{T}\,\nabla_yT \\
J_{Qx} &= L_{12}\,\nabla_x\bar{\mu} + L_{22}\frac{1}{T}\,\nabla_xT + L_{14}\,\nabla_y\bar{\mu} + L_{24}\frac{1}{T}\,\nabla_yT \\
-J_y &= -L_{13}\,\nabla_x\bar{\mu} - L_{14}\frac{1}{T}\,\nabla_xT + L_{11}\,\nabla_y\bar{\mu} + L_{12}\frac{1}{T}\,\nabla_yT \\
J_{Qy} &= -L_{14}\,\nabla_x\bar{\mu} - L_{24}\frac{1}{T}\,\nabla_xT + L_{12}\,\nabla_y\bar{\mu} + L_{22}\frac{1}{T}\,\nabla_yT
\end{aligned}\right\} \quad (7\text{-}195)$$

PROBLEM 7-47 Show that the relations $L_{12} = L_{21}$, $L_{13} = -L_{31}$, etc. are valid by either the Onsager principle or by the fact that Eq. (7-195) must be symmetric with regard to the x- and y-directions.

It is convenient to invert Eq. (7-195) partially as follows:

$$\left.\begin{aligned}
-(1/e)\,\nabla_x\bar{\mu} &= \mathfrak{L}_{11}eJ_x + \mathfrak{L}_{12}eJ_y - \mathfrak{L}_{13}\,\nabla_xT - \mathfrak{L}_{14}\,\nabla_yT \\
-(1/e)\,\nabla_y\bar{\mu} &= -\mathfrak{L}_{12}eJ_x + \mathfrak{L}_{11}eJ_y + \mathfrak{L}_{14}\,\nabla_xT - \mathfrak{L}_{13}\,\nabla_yT \\
J_{Qx} &= -T\mathfrak{L}_{13}eJ_x - T\mathfrak{L}_{14}eJ_y - \mathfrak{L}_{33}\,\nabla_xT - \mathfrak{L}_{34}\,\nabla_yT \\
J_{Qy} &= T\mathfrak{L}_{14}eJ_x - T\mathfrak{L}_{13}eJ_y + \mathfrak{L}_{34}\,\nabla_xT - \mathfrak{L}_{33}\,\nabla_yT
\end{aligned}\right\} \quad (7\text{-}196)$$

where the $\mathfrak{L}_{ij}$ incorporate factors of e and T.

From these equations, we can find the transport coefficients in the following form:

(a) *Isothermal* electric conductivity.

Let $\nabla_xT = \nabla_yT = J_y = 0$

Then $\sigma_i = -eJ_x/(\nabla_x\bar{\mu}/e) = 1/\mathfrak{L}_{11}$

(b) *Adiabatic* electric conductivity.

Let $\nabla_xT = J_{Qy} = J_y = 0$

Then $\sigma a = -eJ_x/(\nabla_x\bar{\mu}/e) = (\mathfrak{L}_{11} - T\mathfrak{L}^2_{14}\mathfrak{L}^{-1}_{33})^{-1}$

(c) Isothermal heat conductivity.

Let $J_x = J_y = \nabla_yT = 0$

Then $\kappa_i = -J_{Qx}/\nabla_xT = \mathfrak{L}_{33}$

(d) Isothermal Hall coefficient.

Let $J_y = \nabla_xT = \nabla_yT = 0$

Then $R_{Hi} = (\nabla_y\bar{\mu}/e)/eBJ_x = \mathfrak{L}_{12}B$

Similar expressions for the remaining coefficients will be found in Callen's paper.[25]

REFERENCES

1. Glicksman, M., "The Magnetoresistivity of Germanium and Silicon," *Progress in Semiconductors*, Vol. 3, ed. A. F. Gibson. New York: John Wiley and Sons, 1958.
2. Bullis, W. M., *Galvanomagnetic Effects in N-Type Germanium*, Tech. Report No. 115. MIT Lincoln Lab., May 14, 1956.
3. Bullis, W. M. and W. E. Krag, *Phys. Rev.*, **101** (1956), 580.
4. Krag, W. E., *Phys. Rev.*, **118** (1960), 435.
5. Pearson, G. L. and C. Herring, *Physica*, **20** (1954) 975.
6. Goldberg, C. and R. E. Davis, *Phys. Rev.*, **94** (1954), 1121.
7. Jones, H., "Theory of Electrical and Thermal Conductivity of Metals," *Encyclopedia of Physics*, Vol. XIX. Berlin: Springer, 1956.
8. Wannier, G. H., *Elements of Solid State Physics.* Cambridge: Cambridge University Press, 1959.
9. Lax, B. and J. G. Mavroides, *Phys. Rev.*, **100** (1955), 1650.
10. Herman, F., *Proc. I.R.E.*, **43** (1955), 1703.

11. Mavroides, J. G. and B. Lax, Lincoln Laboratory Quarterly Progress Report, August 1, 1955 and February 1, 1956.
12. Willardson, R. K., T. C. Harman, and A. C. Beer, *Phys. Rev.*, **96** (1954), 1512.
13. Goldberg, C., E. N. Adams, and R. E. Davis, *Phys. Rev.*, **105** (1957) 865.
14. Long, D., *Phys. Rev.*, **107** (1957) 672.
15. Nussbaum, A. and R. J. Hager, *Phys. Rev.*, **123** (1961), 1958.
16. Seitz, F., *Phys. Rev.*, **79** (1950), 372.
17. Wold, P. I., *Phys. Rev.*, **7** (1916), 169.
18. Blakemore, J. S., G. D. Long, K. C. Nomura, and A. Nussbaum, "Tellurium," *Progress in Semiconductors*, Vol. 6, ed. A. F. Gibson. London: Heywood and Co., 1962.
19. Drabble, J. R., and R. Wolfe, *Proc. Phys. Soc.*, **B69** (1956), 1101.
20. Roth, H., *Measurement of the Galvanomagnetic Coefficients in Pure Tellurium*, Report GA-791. General Atomic Laboratory, 1959.
21. Roth, H., *J. Chem. Phys. Solids*, **8** (1959), 525.
22. Nussbaum, A. and D. Long, *Proceedings of the International Semiconductor Conference.* Prague: Czech Academy of Sciences, 1960.
23. Callen, H. B., *J. Chem. Phys.*, **22** (1954), 518.
24. Long, D., *Phys. Rev.*, **101** (1956), 1256.
25. Callen, H. B., *Phys. Rev.*, **85** (1952), 16.

GENERAL REFERENCES

Semiconductors, ed. N. B. Hannay. New York: Reinhold Publishing Corp., 1959. This collection of review articles extends the material of this chapter. It is also a valuable source of information on many aspects of semiconductor physics.

8

SEMICONDUCTOR PARAMETERS

8-1 Introduction

The purpose of this chapter is to consider some properties of semiconductors which affect the performance of devices. For example, the germanium or silicon used in a transistor must have a specified purity and minority carrier lifetime; we shall discuss the problems involved in specifying these characteristics.

8-2 Resistivity and Compensation

If a semiconductor owed its conduction solely to one kind of carrier — electrons, for example — then a resistivity measurement combined with a knowledge of the mobility would give the purity. From Eq. (7-3), $n = \sigma/e\mu_n$. However, since both electrons and holes are present, we have

$$\begin{aligned}\sigma &= ne\mu_n + pe\mu_p \\ &= \mu_p e(nb + p) \qquad (8\text{-}1)\end{aligned}$$

where $b = \mu_n/\mu_p$. In order to use Eq. (8-1), we must eliminate one of the unknowns n or p; this is done by using

$$np = n_i^2 \qquad (3\text{-}24)$$

This equation is valid for both silicon and germanium at room temperature. Conwell (1958)[1] gives the intrinsic concentrations n_i and corresponding resistivities ρ_i and the mobilities at 300°K as shown in Table 8-1.

TABLE 8-1

	Germanium	*Silicon*
n_i	$2.4 \times 10^{13}/\text{cm}^3$	$1.5 \times 10^{10}/\text{cm}^3$
ρ_i	46 ohm-cm	2.3×10^5 ohm-cm
μ_n	3900 cm²/v-sec	1350 cm²/v-sec
μ_p	1900 cm²/v-sec	480 cm²/v-sec

Putting Eq. (3-24) into Eq. (8-1) gives

$$\sigma = \mu_p e(nb + n_i^2/n) \qquad (8\text{-}2)$$

which is plotted for germanium in Fig. 8-1 and for silicon in Fig. 8-2.

PROBLEM 8-1

(a) Show that the maximum value of σ occurs for the value of n given by

$$n_m = n_i/\sqrt{b} \qquad (8\text{-}3)$$

(b) Show that

$$\sigma_m/\sigma_i = 2\sqrt{b}/(b+1)$$

(c) Show that

$$\rho_m > \rho_i$$

The Hall coefficient, by Eq. (7-19), is

$$R = \frac{-C(nb^2 - p)}{(nb + p)^2} \qquad (8\text{-}4)$$

where C is a constant, and R vanishes when

$$nb^2 = p \qquad (8\text{-}5)$$

Note that R is positive for $n = 0$ and negative for $p = 0$.

PROBLEM 8-2

(a) Show that the Hall coefficient in silicon and germanium is negative for intrinsic material.

(b) Let n_0 denote the value of n for $R = 0$. Show that

$$n_0 = n_i/b \qquad (8\text{-}6)$$

(c) Show that when the resistivity is a maximum

$$R = C(1 - b)/4n_i\sqrt{b} \qquad (8\text{-}7)$$

and is still negative.

(d) Show that the resistivity is

$$\rho_0 = \rho_i \qquad (8\text{-}8)$$

when $R = 0$.

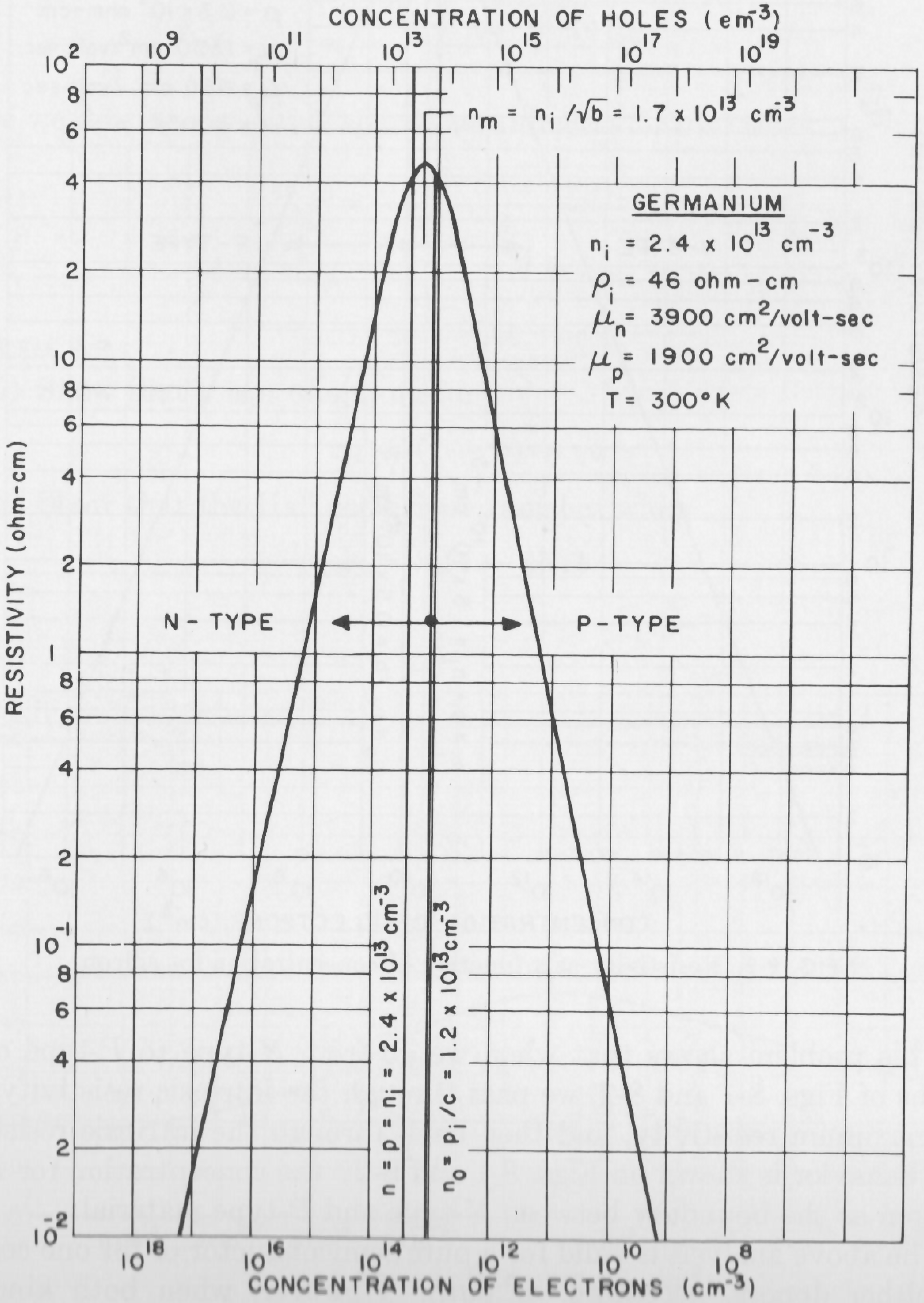


FIG. 8-1. Resistivity as a function of concentration for germanium.

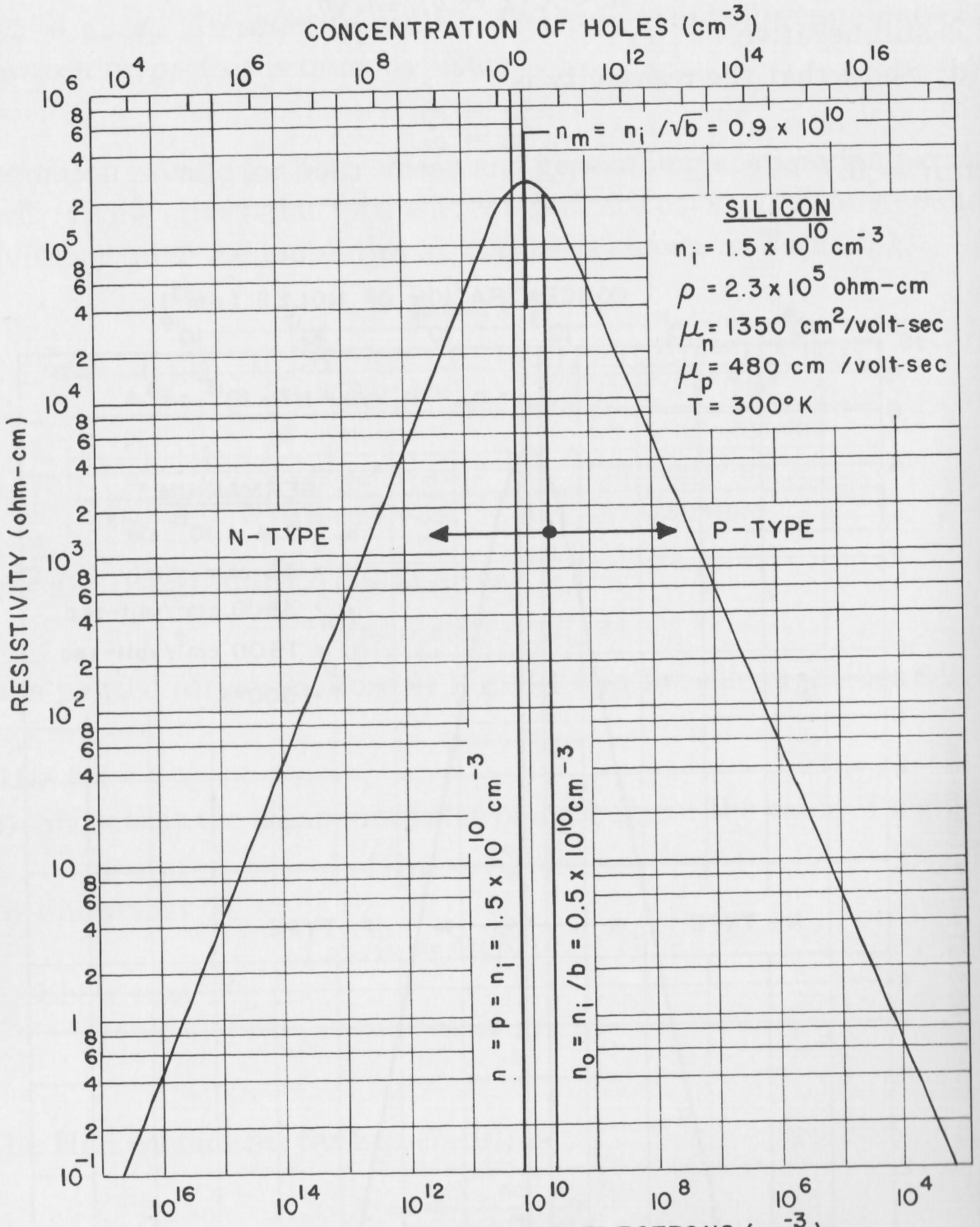


FIG. 8-2. Resistivity as a function of concentration for silicon.

This problem shows that when we go from N-type to P-type on the graphs of Figs. 8-1 and 8-2, we pass through the intrinsic resistivity, then the maximum resistivity, and then back through the intrinsic resistivity. This behavior is shown on Figs. 8-1 and 8-2; the concentration for $R = 0$ is taken as the boundary between N-type and P-type material.

The above analysis is valid for a pure semiconductor or for one containing either donors, acceptors, or both. However, when both kinds are present, we have the condition known as *compensation*, and a resistivity

measurement can not determine the number N_D of donor atoms and N_A of acceptor atoms separately. At room temperature, the impurities in silicon and germanium are ionized, and, by conservation of charge,

$$n - p = N_D - N_A \tag{8-9}$$

Consider an N-type semiconductor and let

$$x = N_D - N_A > 0$$

Then

$$n = p + x = (n_i^2/n) + x$$

or

$$n = \frac{x + \sqrt{x^2 + 4n_i^2}}{2} \tag{8-10}$$

where we drop the negative sign as meaningless. From this

$$\sigma = \frac{Ax + B\sqrt{x^2 + n_i^2}}{2} \tag{8-11}$$

where

$$A = e(\mu_n - \mu_p), \qquad B = e(\mu_n + \mu_p)$$

PROBLEM 8-3

(a) Show that σ is a minimum for

$$x_m = (1 - b)n_i/\sqrt{b}$$

(b) Show that the Hall coefficient vanishes when

$$x_0 = n_i(1 - b^2)/b$$

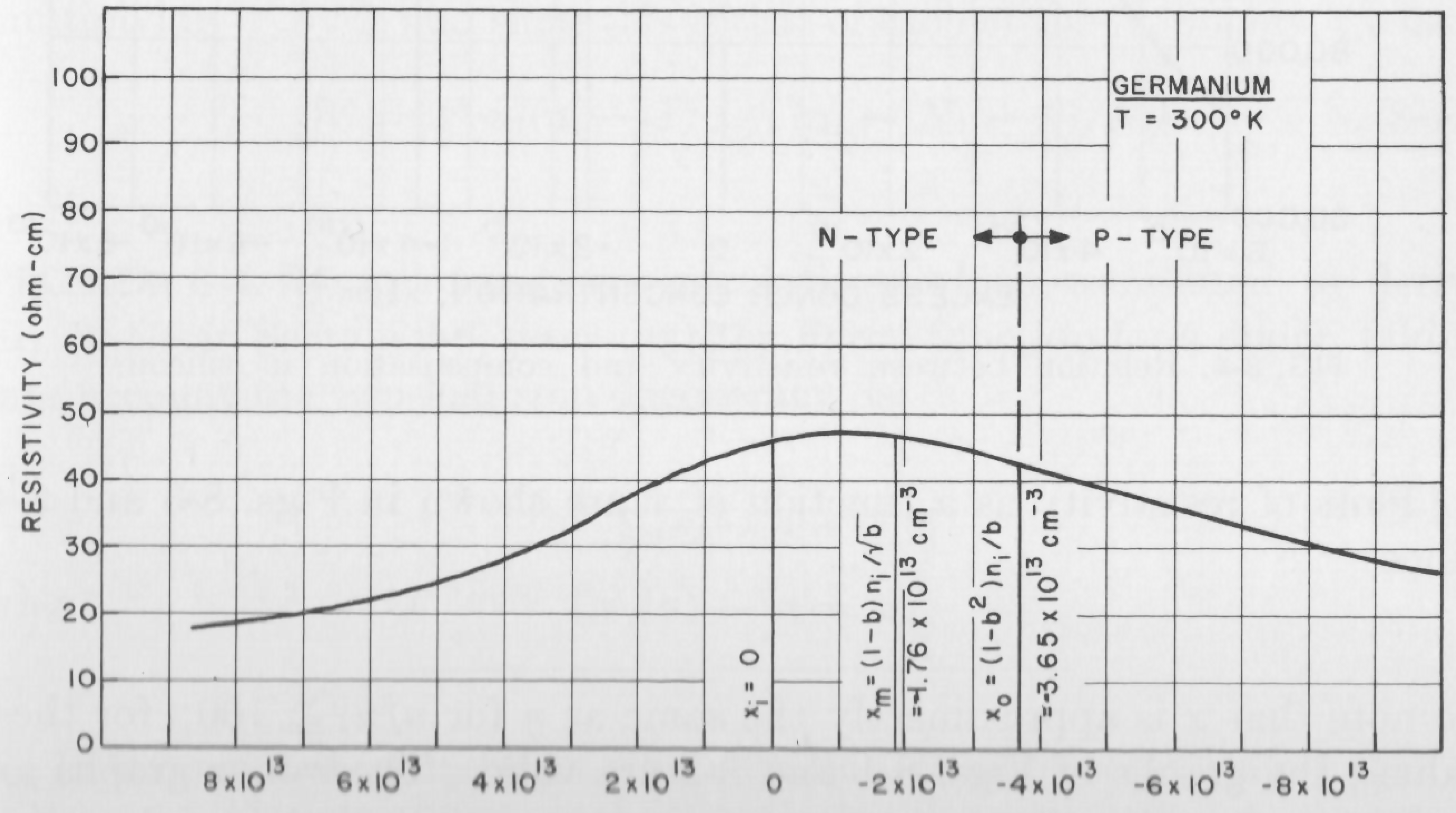


FIG. 8-3. Relation between resistivity and compensation in germanium.

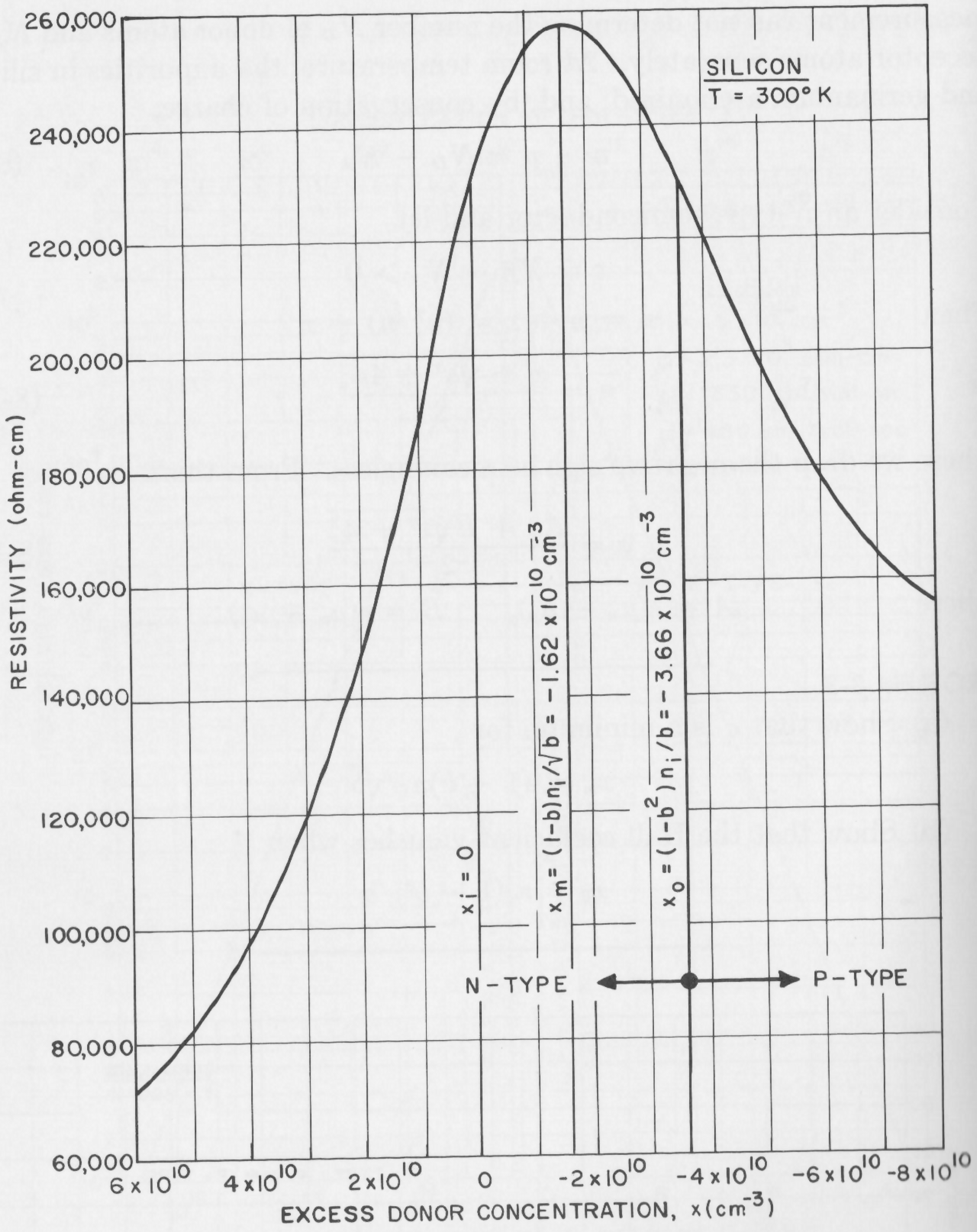


FIG. 8-4. Relation between resistivity and compensation in silicon.

Plots of resistivity as a function of x are shown in Figs. 8-3 and 8-4. Since

$$x = n - (n_i^2/n)$$

we note that x is approximately the same as n for $n/n_i \geqq 100$; for these values, the graphs of Figs. 8-1 and 8-2 are valid. These four graphs are useful in estimating the *excess* impurity concentration, or the total concentration in a lightly compensated material, from the resistivity. However, if

N_D and N_A are increased by equal amounts, x is unchanged, and a resistivity measurement will not detect this. To determine N_D and N_A separately, we must measure the Hall coefficient as a function of temperature. The technique for finding N_D and N_A is the next topic we shall consider.

8-3 Determination of Impurity Concentrations

Before we consider the Hall coefficient as a means of measuring impurity concentrations, we should point out a property of impurity states which we have so far ignored. In the calculation of the density of states in either the conduction or valence bands, the possibility of each level being occupied by two electrons — one with each kind of spin — was included in the theory. For example, in Eq. (7-44), we multiplied the density $1/h^3$ per unit volume of phase space by 2 to incorporate the two-fold spin degeneracy. For donor or acceptor levels, it has been shown by Kohn (1957)[2] that this simple procedure is not sufficient. Consider, for example, an ionized donor which can accept an electron of spin up or spin down, but which afterwards cannot accept either kind without the expenditure of a large amount of energy.

To see what this means quantitatively, let us return to the derivation of the Fermi distribution in Section 2-6. It was shown that the number of ways W of placing N_1 electrons into n_1 states under the Fermi-Dirac rules is

$$W = n_1(n_1 - 1) \ldots (n_1 - N_1 + 1)/N_1!$$

However, any donor can accept an electron with spin up or down, thus multiplying W by 2, and since this is true of each of the N_1 donors, we have

$$W = 2^{N_1} n_1(n_1 - 1) \ldots (n_1 - N_1 + 1)/N_1! \tag{8-12}$$

PROBLEM 8-4 By following the calculations which were used to derive Eq. (2-59) in Section 2-6, show that the Fermi function for a donor, taking into account the two-fold spin degeneracy, is

$$f = \frac{1}{\frac{1}{2}e^{(E_D - E_F)/kT} + 1} \tag{8-13}$$

where E_D is the donor level.

In silicon and germanium, the situation is more complicated than is indicated by Eq. (8-13), since there is further degeneracy due to the structure of the energy bands. In the valence band, for example, we have either four-fold or six-fold degeneracy, including spin, and Eq. (8-13) for acceptor states is modified to

$$f = \frac{1}{\dfrac{e^{(E_A - E_F)/kT}}{g} + 1} \tag{8-14}$$

where g is generally 4 or 6, rather than 2 as above. We shall presently show how it may be determined experimentally.

Let us now consider a compensated N-type semiconductor. By combining Eq. (3-7), (3-11), and (3-13), we obtain

$$n + N_A = \frac{N_D}{\dfrac{e^{(E_F - E_D)/kT}}{g} + 1} \tag{8-15}$$

PROBLEM 8-5

(a) Verify Eq. (8-15).

(b) Eliminate $e^{E_F/kT}$ from Eq. (8-15) and show that it becomes

$$n^2 + n\left[N_A + \frac{N_C}{g}\, e^{-(E_C - E_D)/kT}\right] = \frac{(N_D - N_A)N_C e^{-(E_C - E_D)/kT}}{g} \tag{8-16}$$

A similar expression,

$$p^2 + p\left[N_A + \frac{N_V}{g}\, e^{-(E_V - E_A)/kT}\right] = \frac{(N_D - N_A)N_V e^{-(E_A - E_V)/kT}}{g} \tag{8-17}$$

holds for P-type material.

Some of the unknowns in Eq. (8-16) or Eq. (8-17) can be obtained by going to low temperatures.

PROBLEM 8-6 Show that

$$n = \frac{N_D - N_A}{gN_A}\, N_V e^{-(E_C - E_D)kT} \tag{8-18}$$

when $n \ll N_A$.

Equation (8-18) shows that when the logarithm of the number of free electrons n is plotted against $1/T$, we obtain approximately a straight line whose slope indicates the position of the donor level. The factor $T^{3/2}$ in N_V, as defined by Eq. (3-21), causes the deviation from a straight line. To obtain n (or p), we measure the Hall coefficient at low temperatures, since

$$n \propto 1/Re \tag{7-18}$$

The quantity $(N_D - N_A)$ can be obtained from the curves of the previous section. Figure 8-5 shows a plot of ln ρ and of ln $(1/Re)$ vs $1/T$ for a compensated P-type germanium sample, as measured by Blakemore (1961).[3]

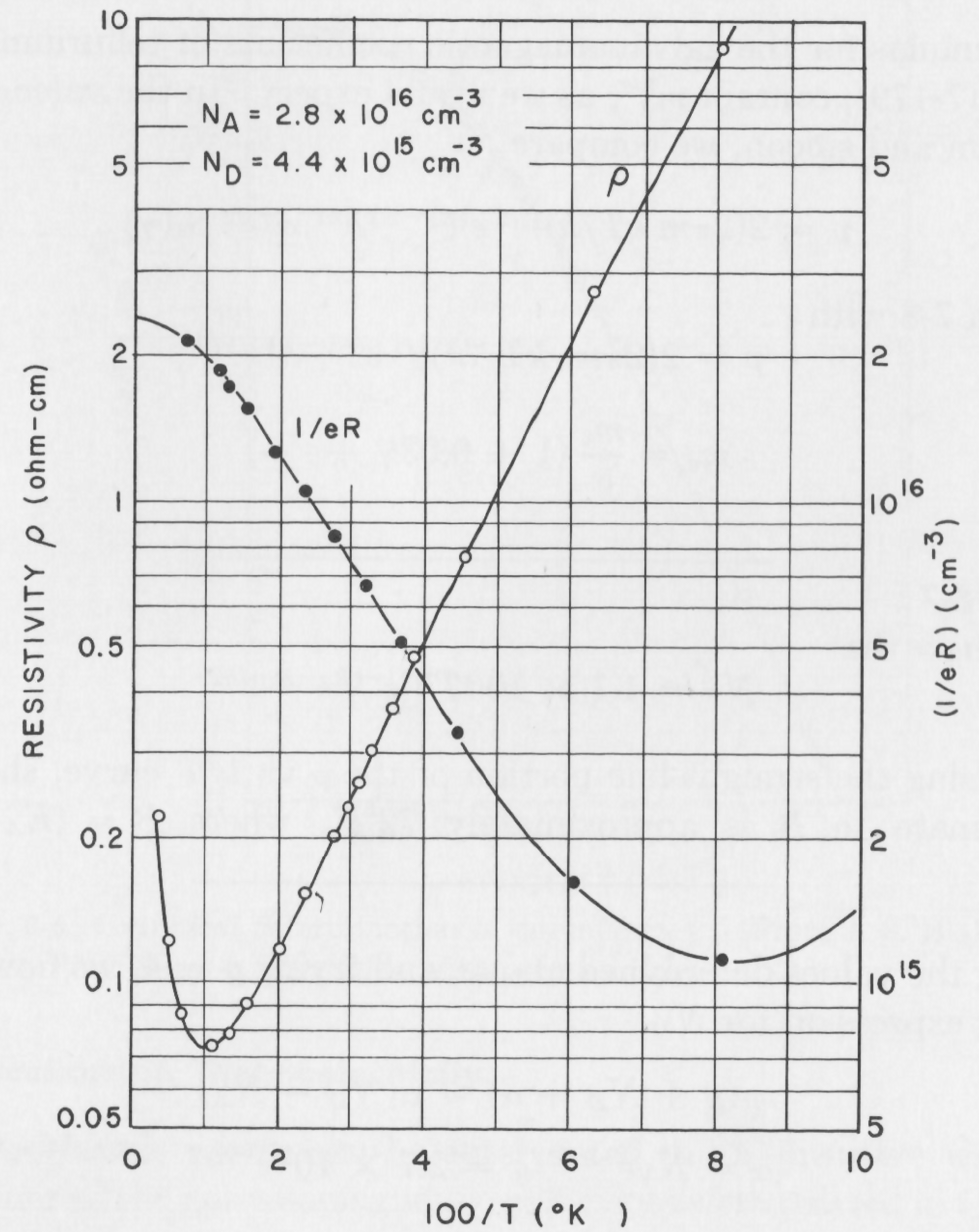


FIG. 8-5. Graph for analysis of Hall coefficient data. (Courtesy of Dr. J. S. Blakemore, Honeywell Research Center.)

The value of $(N_D - N_A)$ also comes from the limiting value of the curve of $(1/Re)$ as T approaches infinity, for the flattening of this curve at the left-hand end represents the *exhaustion region.* This corresponds to temperatures for which the germanium has been warmed up enough for all the impurities to be fully ionized, or exhausted, but for which intrinsic conduction is not yet significant. Hence, the Hall coefficient approaches a value corresponding to $(N_D - N_A)$.

The rest of the procedure for finding N_A involves some complicated curve-fitting. The value of the effective mass for holes, as taken from Lax and Mavroides (1955)[4], is about 0.37. This value, used to fit Hall effect data, is called the *density-of-states effective mass* m_d. For a single ellipsoid, it is defined as the mean of the three principal masses.

$$m_d = (m_1 m_2 m_3)^{1/3}$$

so the formulas for the galvanomagnetic coefficients of tellurium, Eqs. (7-178) and (7-179), contain $m_d^{3/2}$, as we would expect. In the valence band of germanium and silicon, we compare

$$p = 2(2\pi m\, kT/h^3)^{3/2}\, e^{(E_V - E_F)/kT}\, \alpha^{-3/2}\, S_0(\gamma)$$

of Section 7-8 with

$$p = 2(2\pi m_d\, kT/h^3)^{3/2}\, e^{(E_V - E_F)/kT} \tag{3-23}$$

to obtain

$$m_d = \frac{m_0}{\alpha}(1 + 0.03\gamma + \cdots)$$

PROBLEM 8-7

(a) Show that

$$N_V = 1.1 \times 10^{14} T^{3/2} e^{-(E_A - E_C)/kT} \tag{8-19}$$

(b) Using the straight-line portion of the ρ vs $1/T$ curve, show that a first estimate to B is approximately 72°K, where $B = (E_A - E_C)/k$.

Using the values determined above, and trying $g = 4$, we now have the following expression for N_D:

$$p(p + N_D + \alpha) = \alpha(N_D - N_A) \tag{8-20}$$

where

$$\alpha = N_V e^{-B/T}/g = 2.7 \times 10^{14} e^{-72/T}$$

PROBLEM 8-8

(a) Choosing a temperature well down on the $1/Re$ curve, for example, $T = 25$ deg, show that an approximate solution to Eq. (8-20) is $N_D = 4.5 \times 10^{15}$.

(b) Plot p as a function of $1/T$ and determine what adjustments have to be made for a better fit.

The method outlined above has been used to determine g experimentally For example, Blakemore (1959)[5] has given the curve of Fig. 8-6 for a P-type germanium sample with 2.7×10^{15} indium atoms/cm^3 and 4.8×10^{14} antimony atoms/cm^3. It is seen from this data that, with careful experimental methods, it is possible to distinguish between $g = 2$ and $g = 4$.

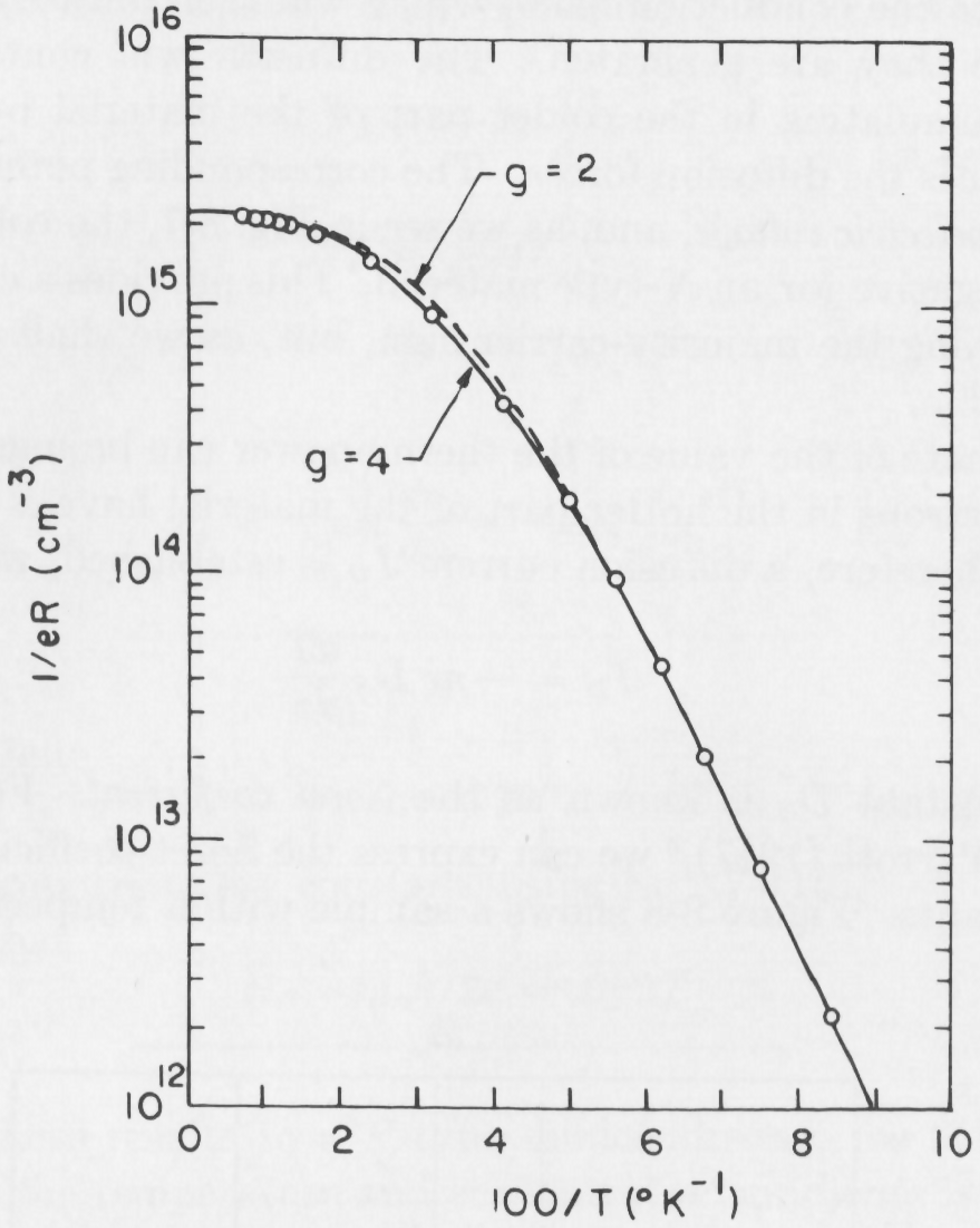


FIG. 8-6. Graphical determination of degeneracy g. (From J. S. Blakemore, Phil. Mag. **4**, 560, 1959, with permission.)

8-4 Applications in Thermoelectricity

The thermoelectric power can be interpreted in an intuitive way in the same manner as the galvanomagnetic coefficients were treated in Chapter 7. Consider a block of N-type semiconductor which is heated in one spot, as shown in Fig. 8-7. This local heating will excite an excess number of

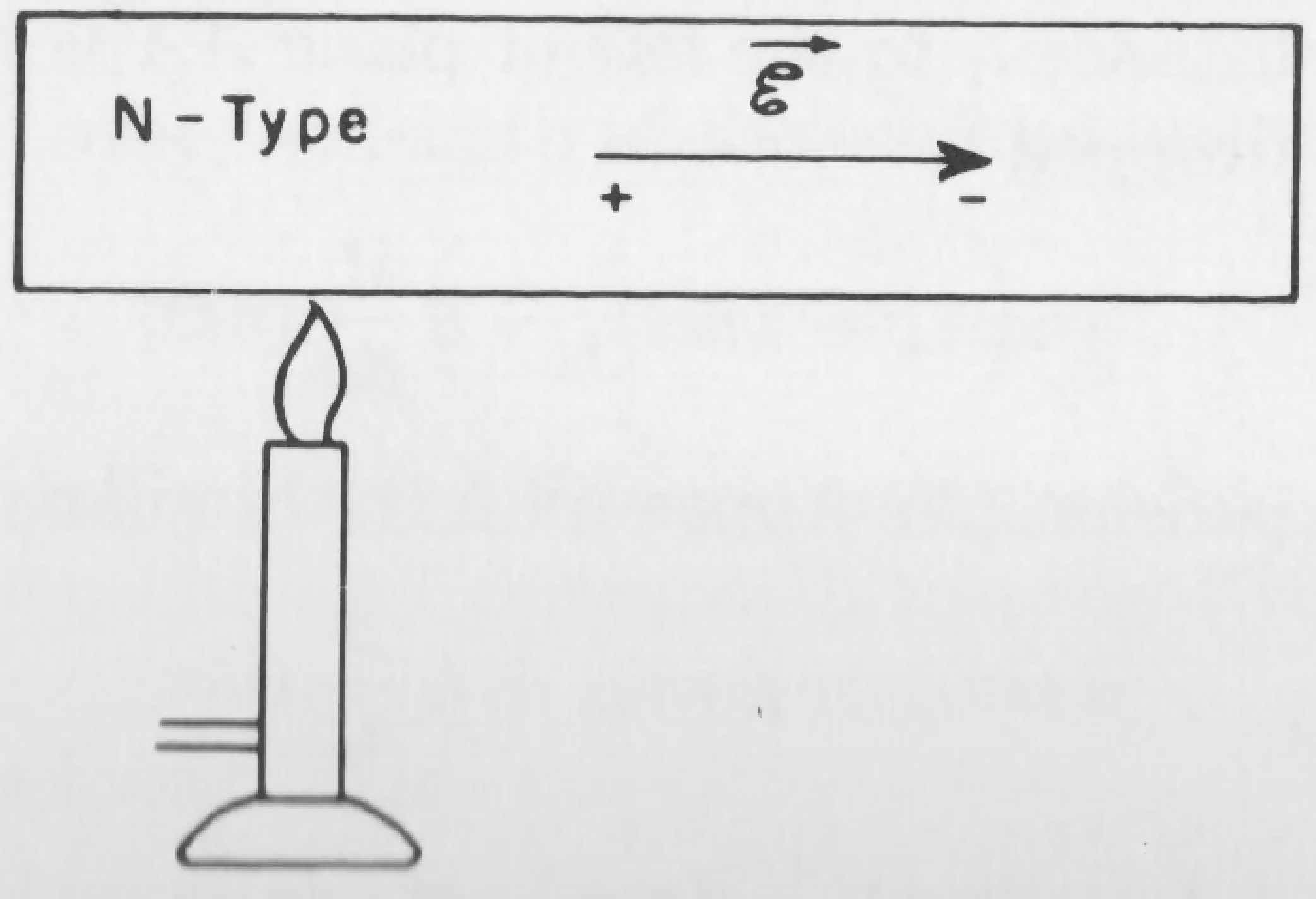


FIG. 8-7. Principle of hot-probe type determination.

electrons into the conduction band, which will then diffuse away from the point where they are generated. The diffusion will continue until the charges accumulating in the colder part of the material build up a field which balances the diffusion forces. The corresponding potential difference is the *thermoelectric voltage*, and, as we see in Fig. 8-7, the colder part of the sample is negative for an N-type material. This provides a convenient way for determining the majority-carrier sign, but, as we shall show, it is not always reliable.

An estimate of the value of the thermopower can be made on the basis that the electrons in the hotter part of the material have a higher average velocity. Therefore, a diffusion current J_D is established, where

$$J_D = -ne\,D_S \frac{dT}{dx} \tag{8-21}$$

and the constant D_S is known as the *Soret coefficient*. Following Rose, Billig, and Parrott (1957),[6] we can express the Soret coefficient in terms of other constants. Figure 8-8 shows a sample with a temperature gradient;

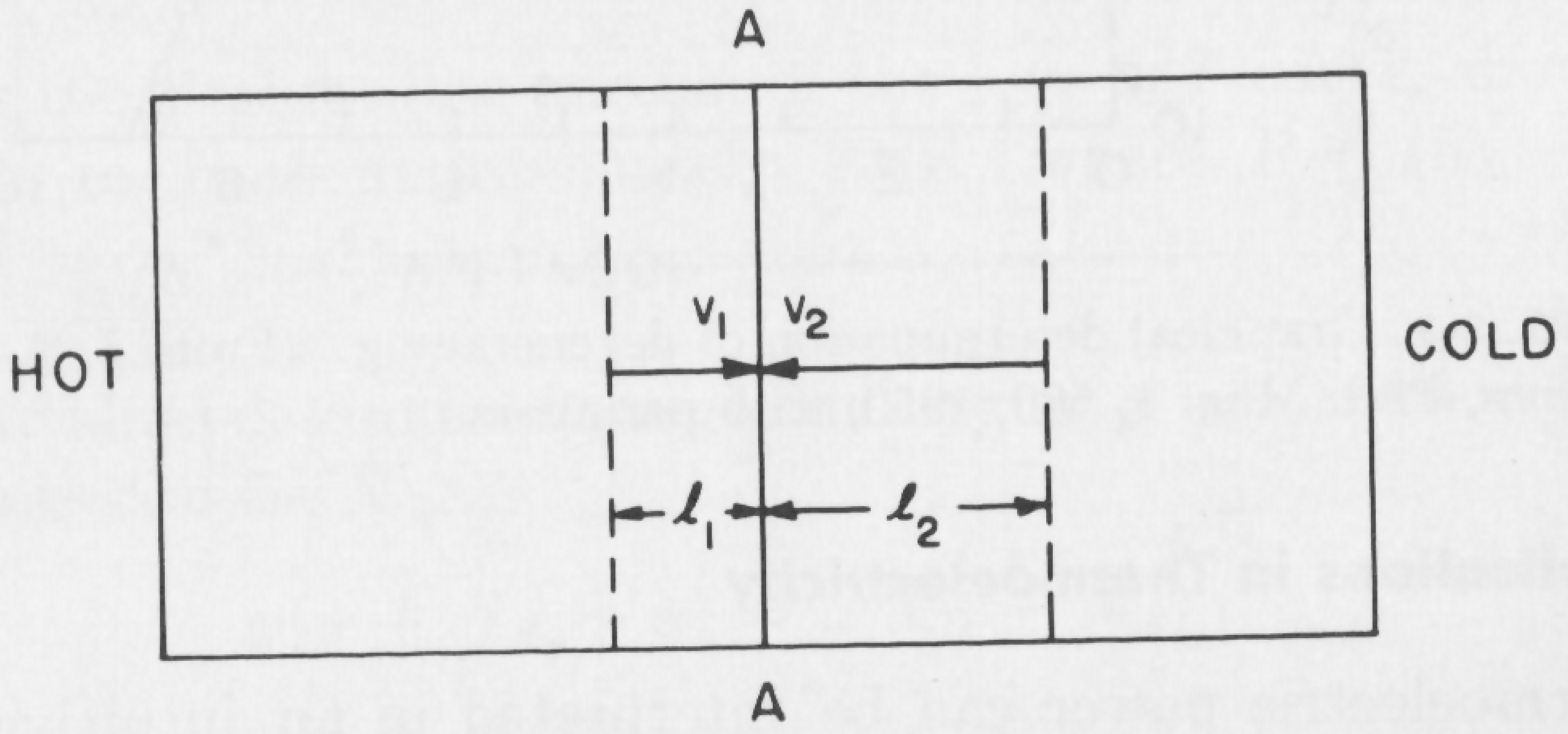


FIG. 8-8. Diagram used to discuss Soret effect. (See text.)

we shall assume that the thermal motion of the charge carriers takes place along the x-axis. If v denotes the thermal velocity of the charge carriers, the current at a distance l_1 to the left of plane AA is $J_x = (nev)_1$. At the plane AA, J_x has dropped to

$$(nev)_A = (nev)_1 - l_1 \frac{d}{dx}(nev)$$

using a Taylor expansion (*see* Appendix A1). Similarly,

$$(nev)_A = (nev)_2 + l_2 \frac{d}{dx}(nev)$$

Hence,
$$J_x = (nev)_1 - (nev_2) = -le \frac{d}{dx}(nv)$$

where l is the mean free path. If v is constant,

$$J_x = -lev \frac{dn}{dx}$$

But
$$J_D = -eD \frac{dn}{dx} \tag{3-37}$$

so
$$D = lv$$

If n is constant,

$$J_x = -nel \frac{dv}{dx} = -nel \frac{dv}{dT} \frac{dT}{dx}$$

PROBLEM 8-9

(a) Show that

$$v = CT^{1/2}$$

where C is a proportionality constant, using Eq. (7-115).

(b) Show that

$$D_S = \mu_n k/2e = D/2T \tag{8-22}$$

Applying these results to a P-type semiconductor, the total diffusion current due to the temperature and concentration gradients is

$$J_D = -eD \frac{dp}{dx} - ep\,D_S \frac{dT}{dx} \tag{8-23}$$

For the situation of Fig. 8-7, this must be balanced against the electric current.

PROBLEM 8-10 Show that the above condition leads to

$$\frac{d\mathcal{V}}{dx} = -\frac{k}{e}\left(T \frac{d}{dT} \ln p + \tfrac{1}{2}\right) \frac{dT}{dx} \tag{8-24}$$

Integrating this over the length of the semiconductor gives

$$\Delta\mathcal{V} = \frac{k}{e}\left(T \frac{d}{dT} \ln p + \tfrac{1}{2}\right) \Delta T \tag{8-25}$$

where $\Delta\mathcal{V}$ is the small voltage difference corresponding to a temperature difference ΔT, and a positive ΔT corresponds to a negative $\Delta\mathcal{V}$ in accordance with the sign conventions we have previously used.

The chemical potential K is

$$K = kT \ln \frac{n_i}{p} \tag{3-35}$$

and since $N_V \propto T^{3/2}$, and $(E_V - E_F)$ is independent of x, we have

$$\Delta K = -k\left(\ln\frac{N_V}{p} + \frac{3}{2}\right) + \frac{kT}{e}\frac{d}{dT}(\ln p)\,\Delta T$$

The change in the electrochemical potential corresponding to ΔK and $\Delta\mathcal{V}$ is then

$$\begin{aligned}\Delta\bar{\mu} &= \Delta K - e\Delta\mathcal{V}\\ &= -k\left(\ln\frac{N_V}{p} + 2\right)\Delta T\end{aligned}$$

From Eq. (6-36), the thermoelectric power is

$$\epsilon = \frac{k}{e}\left(\ln\frac{N_V}{p} + 2\right) \tag{8-26}$$

By Eq. (3-23), this becomes

$$\epsilon = \frac{k}{e}\left(\frac{E_F - E_V}{kT} + 2\right) \tag{8-27}$$

Since the Fermi level is generally a few kT above the top of the valence band, this formula predicts that the thermopower in a semiconductor will be of the order of a few hundred microvolts/deg, where $k/e = 86\ \mu\text{V/deg}$.

PROBLEM 8-11

(a) Show that the thermopower for a semiconductor with comparable electron and hole densities is

$$\epsilon = \frac{-k}{e(bn + p)}\left[bn\left(\ln\frac{N_C}{n}\right) - p\left(\ln\frac{N_V}{p}\right) + 2(bn - p)\right] \tag{8-28}$$

where b is the mobility ratio.

(b) Show from Eq. (8-28) that the thermoelectricity of a P-type semiconductor can change sign from positive to negative as the temperature is increased. Show that this reversal takes place when the relation

$$p = bn$$

is approximately true, provided b is not too large. Compare this with the condition for the Hall coefficient to reverse.

The transport theory of Chapter 7 can be combined with some of the results of Chapter 6 to study thermoelectric generators as possible devices for the direct conversion of heat into electricity. The definition of absolute thermopower was given as

$$\epsilon = -\operatorname{grad}\bar{\mu}/e\operatorname{grad}T)_{J=0} \tag{6-36}$$

and for $J = 0$, Eq. (7-59) becomes

$$0 = (e\mathcal{E}_i - \partial E_F/\partial i)I_{11} + [(E_F/T)I_{11} + (I_{21}/T)]\,\partial T/\partial i \qquad (i = x, y, z) \tag{8-29}$$

where two subscripts are used on the transport integrals, as follows:

$$I_{l1} = \int \tau E^{l+1/2}\,\partial f_0/\partial E \tag{7-58}$$

$$I_{1n} = \int \tau^n E^{3/2}\,\partial f_0/\partial E/\Delta \tag{7-83}$$

The Fermi level E_F enters Eq. (8-29) via the Fermi function f_0; the derivation of this distribution was established by considering uncharged particles. Hence, we identify E_F with K, so the first term of Eq. (8-29) will be recognized as

$$(e\mathcal{E}_i - \partial K/\partial i) = -\partial\bar{\mu}/\partial i \tag{8-30}$$

Combining this with the definition of ϵ, Eqs. (6-36) and (8-29), we have

$$e\epsilon T = (I_{21}/I_{11}) - E_F \tag{8-31}$$

This equation has an interpretation which is similar to those of the equations we discussed in Section 7-1. By Eq. (6-45), the left-hand side is proportional to the Peltier heat Π, and by using Eq. (7-124), the right-hand side can be transformed to give

$$e\epsilon T = e\Pi = <(E - E_F)\tau>/<\tau> \tag{8-32}$$

PROBLEM 8-12

(a) Give the definition of $< E_T >$ used in Eq, (8-32).

(b) Show that if $(E - E_F) \gg kT$, then Eq. (8-32) simplifies to

$$\epsilon = (E - E_F)/eT \tag{8-33}$$

thus showing why the thermopower is often used as a direct means of detecting changes in the Fermi energy.

(c) Consider the electrons in the conduction band of a semiconductor. By comparing the value of $(E - E_F)$ for these electrons with the conduction electrons in a metal, show why we would expect a semiconductor to have a higher value of ϵ.

The thermal conductivity κ can be calculated from Eq. (7-61). By Eq. (6-34), κ is the ratio of $-J_Q$ to $\partial T/\partial i$ for $J = 0$.

PROBLEM 8-13

(a) Show that

$$\kappa = (I_{11}I_{31} - I_{21}^2)/I_{11}T \tag{8-34}$$

(b) Use the definitions of I_{nl} and $< \tau^n >$ to show that Eq. (8-31) reduces to Eq. (8-27) for lattice scattering.

We will now give a graphical interpretation of Eq. (8-31), due to Domenicali (1958),[7] which enables us to get a physical picture of the various factors which contribute to the thermoelectric power. This equation can be arranged as follows:

$$e\epsilon T = \frac{\int_0^\infty (E - E_F)E^{3/2}\tau(E)\ \partial f_0/\partial E\ dE}{\int_0^\infty E^{3/2}\tau(E)\ \partial f_0/\partial E\ dE} \tag{8-35}$$

If we write $E^{3/2}\ \tau(E)\ \partial f_0/\partial E$ as a single function $F(E)$, we see that Eq. (8-35) becomes

$$e\epsilon T = \int F(E)(E - E_F)\ dE \Bigg/ \int F(E)\ dE \tag{8-36}$$

This equation is of the same form as the definition of the *center-of-mass* of an arbitrary body. Consider, as an example, two particles with masses m_1 and m_2. The center-of-mass $\bar{x}$ of this group of two particles is defined as

$$\bar{x} = \frac{m_1x_1 + m_2x_2}{m_1 + m_2} \tag{8-37}$$

where x_1 and x_2 are the coordinates of the two particles. For a solid body, Eq. (8-37) is generalized to

$$\bar{x} = \int x\ dm \Big/ \int dm \tag{8-38}$$

Defining the *density* ρ as the mass per unit volume, Eq. (8-38) can be expressed

$$\bar{x} = \int x\rho(x)\ dV \Big/ \int \rho(x)\ dV \tag{8-39}$$

where ρ may be a function of position; Eq. (8-39) is seen to be of the same form as Eq. (8-36). This analogy can be seen on the graph of Fig. 8-9; from Problem 7-30, we know that $\partial f_0/\partial E$ is a pulse about 2 kT wide. We have plotted $\partial f_0/\partial E$ and $E^{3/2}$ on the graph from a knowledge of their functional form. Although $\tau(E)$ is also shown, its behavior represents a guess. We also show as a dotted curve the product of $F(E) = \tau(E)E^{3/2}\ \partial f_0/\partial E$. We see that the center-of-mass lies slightly to the right of E_F and that a way in which we can hope to move the center-of-mass farther out, and hence get a higher thermopower, is to find a material whose τ vs E curve has the odd shape shown in Fig. 8-10.

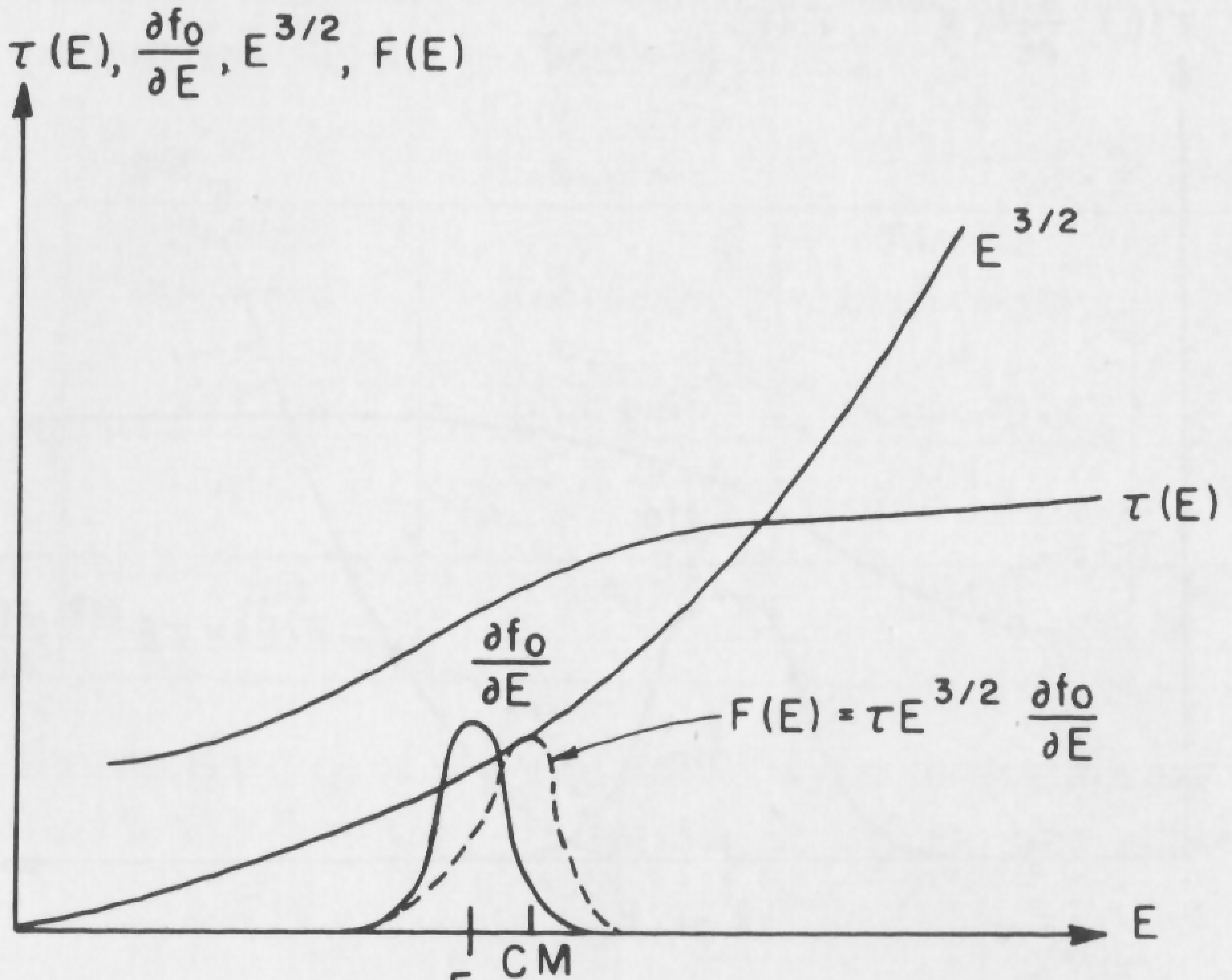


FIG. 8-9. Graphical interpretation of thermoelectric power. (From C. A. Domenicali, Phys. Rev. **112,** 1863, 1958, with permission.)

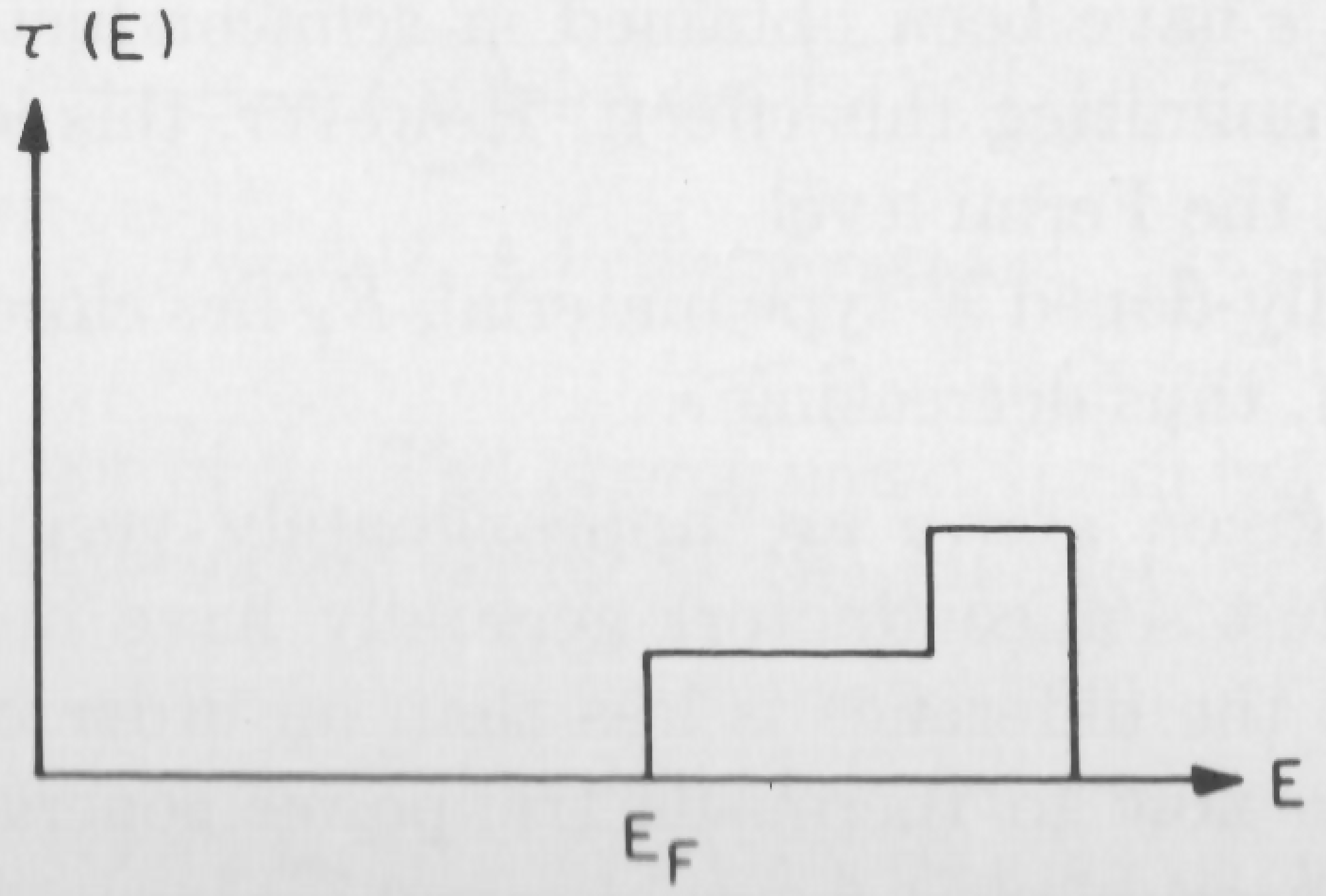


FIG. 8-10. Shape of relaxation time *vs.* energy curve which would give a high thermopower.

It should be realized that the arguments presented so far apply to metals. To appreciate this, we show in Fig. 8-11 the analogous situation for semiconductors. As we discussed in Chapter 3, for an N-type material we take the zero of kinetic energy as the bottom of the conduction band, denoted by E_c. Hence, the curve representing the product function F lies wholly to the right of E_F, and we would expect a higher thermopower than we do for metals. There are, however, two effects which tend to limit this apparent superiority of metals over semiconductors. These are

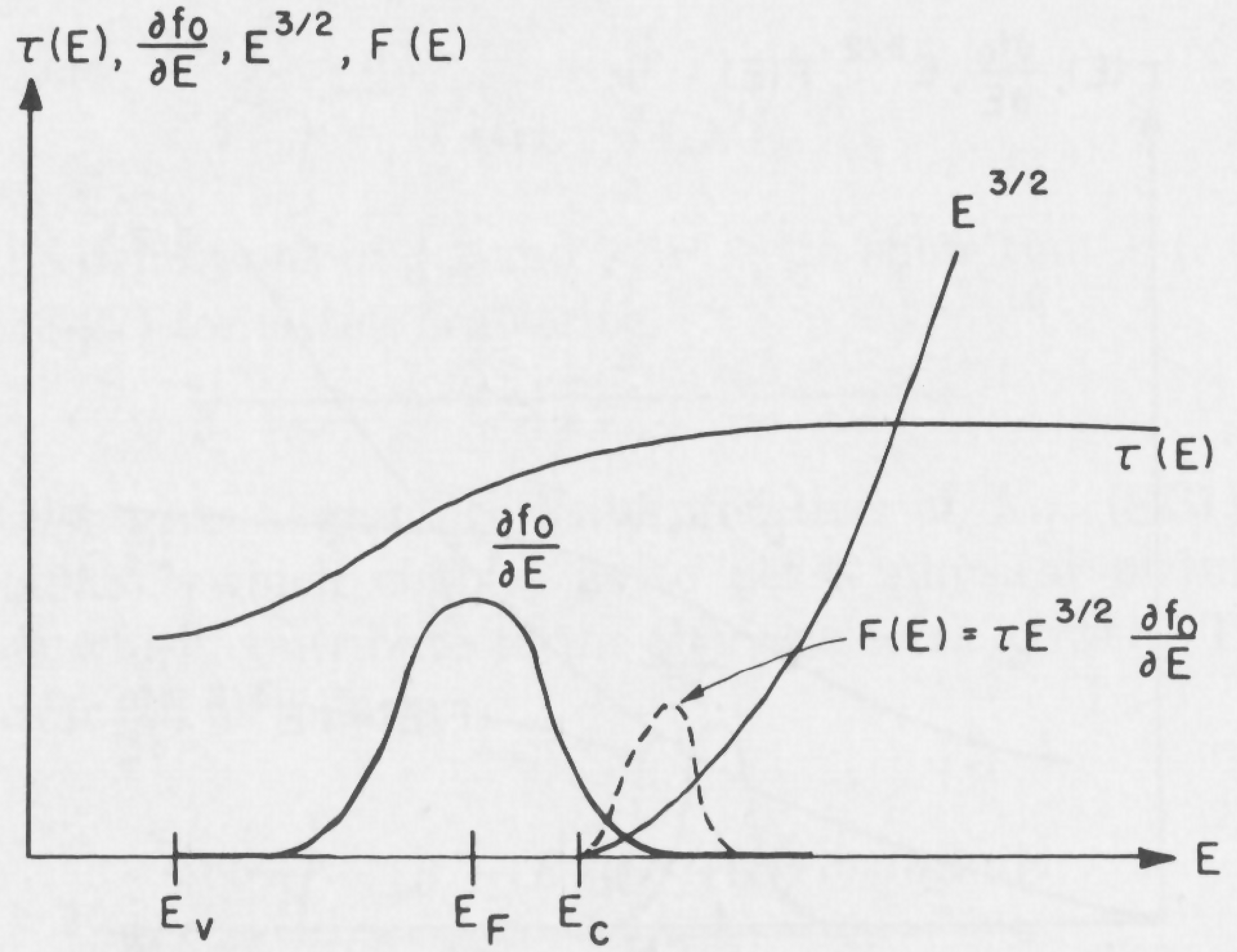


FIG. 8-11. Graphical interpretation of thermoelectricity as applied to a semiconductor.

(1) The contribution due to holes
In assuming an N-type material, we have ignored the effect of holes, which will add a small area to the left of E_V, the top of the valence band, thus shifting the center-of-mass back towards E_F. High values of ϵ have been obtained in semiconductors by very heavy doping, minimizing this effect. However, this leads to:

(2) A shift in the Fermi level
In a heavily-doped N-type material, E_F lies close to E_c (as shown in Fig. 3-11), thus decreasing ϵ.

The arguments given above are approximately verified experimentally, where we find that semiconductors generally have higher thermopowers than metals, but the difference is less than an order of magnitude.

We shall turn now to thermoelectric power generators. Joffe (1960)[8] has defined the *thermoelectric figure of merit* Z as

$$Z = \epsilon^2/\kappa\rho \tag{8-40}$$

where κ is the thermal conductivity and ρ is the electrical resistivity. In this formula, ϵ^2/ρ represents the power being generated by the thermoelectric material, and κ is a measure of its ability to sustain a temperature difference, thus explaining the form of the definition. One of the most efficient materials for this purpose is bismuth telluride, Bi_2Te_3, which has the physical properties shown in Table 8-2.

TABLE 8-2

	Seebeck coefficient ϵ (μV/°K)	*Resistivity* ρ (ohm-cm)	*Thermal conductivity* κ (w/cm-°K)	*Figure of Merit* Z (°K^{-1})
N-type (at 100°C)	−210	1.05×10^{-3}	0.016	2.8×10^{-3}
P-type (at 25°C)	+195	9.0×10^{-4}	0.015	2.6×10^{-3}

Since the thermopowers of N-type and P-type materials are opposite in sign, we can use a PN-junction to double, in effect, the value of ϵ or to multiply Z by a factor of 4. The arrangement for doing this is shown in Fig. 8-12. The thermoelectric power of the junction, from Eq. (8-31), is

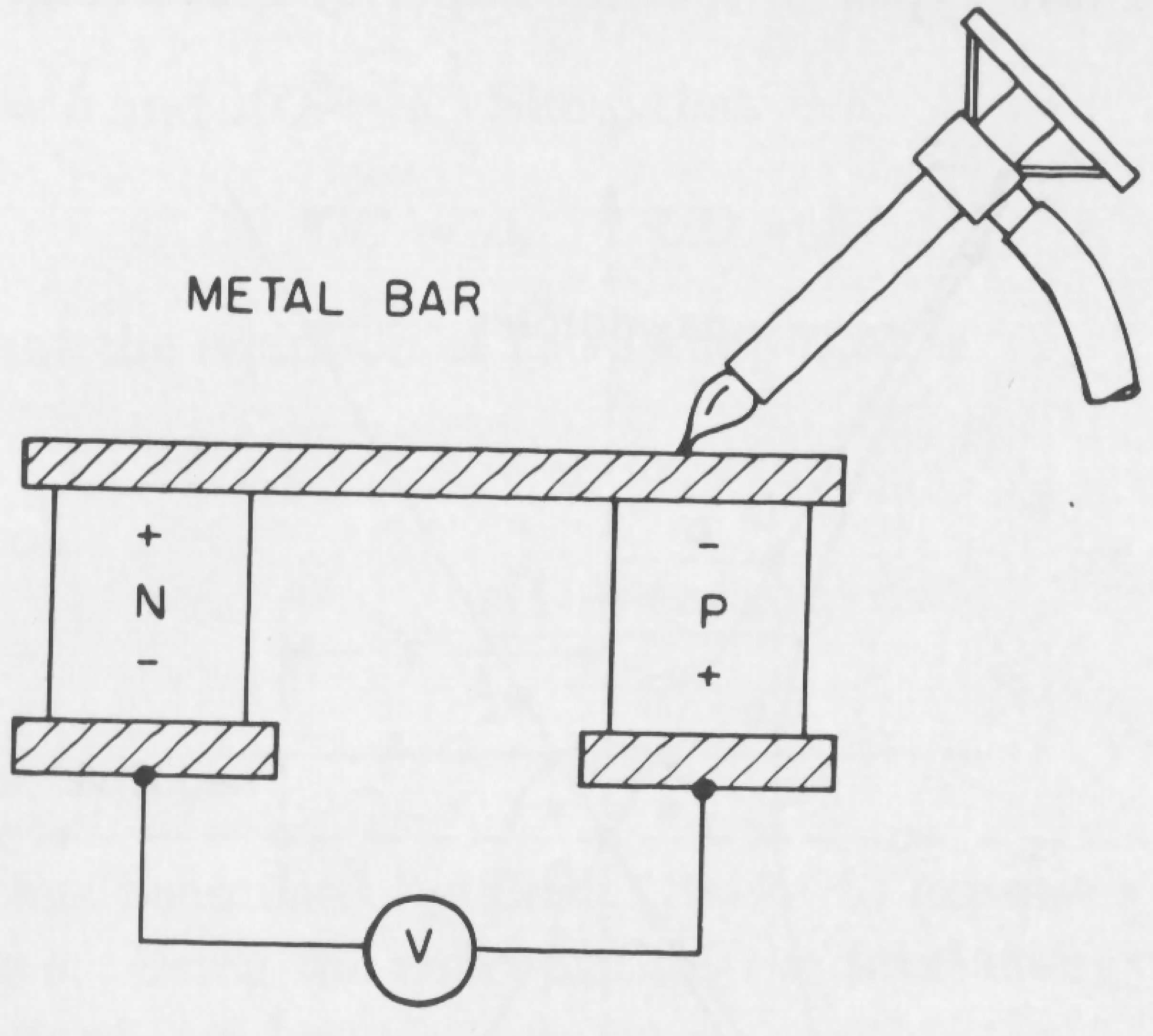


FIG. 8-12. A Peltier generator.

$$\begin{aligned}\epsilon_{PN} &= \epsilon_P + \epsilon_N \\ &= \frac{1}{eT}\left[\left(\frac{I_{21}}{I_{11}}\right)_P + \left(\frac{I_{21}}{I_{11}}\right)_N - (E_{FP} + E_{FN})\right]\end{aligned} \tag{8-41}$$

PROBLEM 8-14 By recalling that the Fermi level enters the theory of thermoelectricity via the term $\partial f_0/\partial E$, show that $(E_{FP} + E_{FN})$ in Eq. (8-41) equals the gap width.

8-5 Scattering and Mobility

The mobility of electrons and holes is one of the important properties of a semiconductor, as can be seen by reviewing the derivations in Chapter 4. In our discussion of diodes and transistors, we assumed that the operating temperature for these devices was approximately 300°K. However, most transistors will rise in temperature when the current density rises above a small fraction of their rated values, and we now wish to consider the variation of mobility and the associated electrical properties as a function of temperature.

The approach to this problem is through a study of *scattering*, defined as the process which causes an electron or hole to deviate from its normal path. Scattering is caused by a number of mechanisms, but we shall consider only a few of them here from an elementary viewpoint. One of the most important types is *ionized impurity* scattering, which can be regarded as the interaction of an electron or hole with an occupied, or ionized, donor or acceptor in accordance with Coulomb's inverse square law, Eq. (1-12). Regarding this as a problem in mechanics, we use the fact that astronomical bodies move in orbits which are elliptic, parabolic, or hyperbolic, depending upon the intitial velocity and angular momentum.

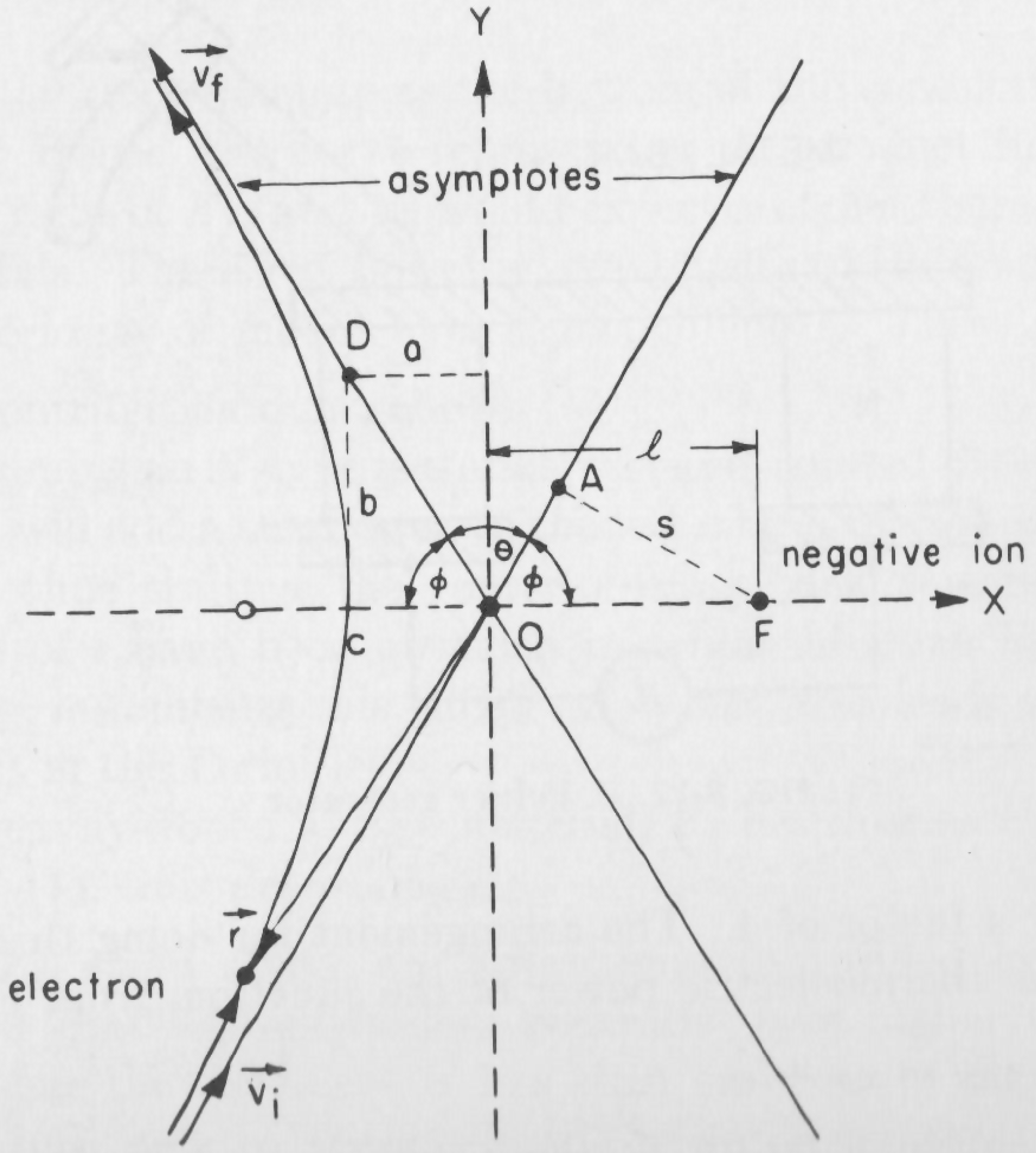


FIG. 8-13. Diagram used in derivation of Rutherford scattering law. (From M. Born, *Atomic Physics*, Blackie & Son, Glasgow, 1944, with permission.)

Let us assume that an electron approaching a negatively charged acceptor is scattered in the hyperbolic path of Fig. 8-13. The two lines through the origin making an angle of θ are called the *asymptotes of the* hyperbola and represent the directions of the initial velocity $\mathbf{v}_i$ and the final velocity $\mathbf{v}_f$. The distance s from the location of the scattering ion to the original direction of motion is called the *impact parameter*.

PROBLEM 8-15

(a) Center a set of axes XOY on a sheet of graph paper and place a scattering center at a point F on the positive x-axis, as shown in Fig. 8-13. Plot the left branch of the hyperbola

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1$$

using $AF = s = b$ and $AO = a$. Show that

$$OC = a, \qquad CD = b$$

(b) Show that the equation of the asymptotes is

$$y = \pm bx/a$$

(c) Show that

$$l = \frac{s(1 + \cos\varphi)}{\sin\varphi} \tag{8-42}$$

This result has been used by Born (1944)[9] to express s in terms of the scattering angle θ. Using the concept that the total energy of the electron must be a constant, we have

$$\tfrac{1}{2}mv_i^2 = \tfrac{1}{2}mv_0^2 + \frac{e^2}{4\pi\epsilon l} \tag{8-43}$$

where v_0 is the velocity of the electron at the point C. Solving for $(v_0/v_i)^2$ yields

$$\left(\frac{v_0}{v_i}\right)^2 = 1 - \frac{2k}{s}\frac{\sin\varphi}{1 + \cos\varphi} \tag{8-44}$$

where $k = e^2/4\pi\epsilon mv_i^2$.

The angular momentum $\mathbf{p}_\theta$ of the electron is also conserved, for the repulsive force can be written

$$\mathbf{F} = m\frac{d^2\mathbf{r}}{dt^2}$$

so that

$$\mathbf{r} \times \mathbf{F} = m\mathbf{r} \times \frac{d^2\mathbf{r}}{dt^2} = 0$$

since $\mathbf{r}$ and $\mathbf{F}$ are colinear. Now

$$\frac{d}{dt}m\mathbf{r} \times \frac{d\mathbf{r}}{dt} = m\mathbf{r} \times \frac{d^2\mathbf{r}}{dt^2} + m\frac{d\mathbf{r}}{dt} \times \frac{d\mathbf{r}}{dt}$$

The two terms on the right vanish, and the left-hand side integrates to

$$m\mathbf{r} \times \frac{d\mathbf{r}}{dt} = C$$

By definition [Eq. (1-12)], we have that $\mathbf{p}_\theta$ is a constant. From this, we see that

$$msv_i = mlv_0 \qquad (8\text{-}45)$$

or

$$\frac{v_0}{v_i} = \frac{s}{l} = \frac{\sin\varphi}{1 + \cos\varphi}$$

and

$$\left(\frac{v_0}{v_i}\right)^2 = \frac{\sin^2\varphi}{(1 + \cos\varphi)^2} = \frac{1 - \cos\varphi}{1 + \cos\varphi} \qquad (8\text{-}46)$$

Combining Eqs. (8-44) and (8-46) gives

$$\frac{1 - \cos\varphi}{1 + \cos\varphi} = 1 - \frac{2k}{s}\,\frac{\sin\varphi}{1 + \cos\varphi}$$

from which

$$s = k\tan\varphi$$

Since $\theta = \pi - 2\varphi$, this becomes

$$s = k\cot\theta/2 \qquad (8\text{-}47)$$

which is one form of the *Rutherford scattering law.*

The calculations performed so far have been based on regarding the electron as a classical particle, obeying the ordinary laws of mechanics. It is possible to obtain similar results by treating the electron as a wave and solving the Schroedinger equation. The solution is accomplished by a method known as the *Born approximation,* which is based on the idea that the incident and scattered waves can be regarded as non-interacting. (A simple discussion of this method will be found in Born's book.[9]) The wave approach is also necessary at this point in order to differentiate between the relaxation time τ used so extensively in Chapter 7 and the *mean free time between collisions* τ_c, as defined by Shockley (1950).[10]

If W_{if} is the probability per unit time and per unit volume of momentum-space that an electron with intial velocity $\mathbf{v}_i$ suffers a collision resulting in a velocity $\mathbf{v}_f$, then the current associated with this transition is

$$d\mathbf{J} = dt\int e(\mathbf{v}_f - \mathbf{v}_i)nW_{if}\,dV \qquad (8\text{-}48)$$

where $dV = (mv_i)^2\sin\theta\,d\theta\,d\varphi$. If we choose $\mathbf{v}_i$ along the z-axis of this space, then

$$\mathbf{v}_i = \mathbf{k}v_i$$

$$\mathbf{v}_f = (\mathbf{i}\sin\theta\cos\varphi + \mathbf{j}\sin\theta\sin\varphi + \mathbf{k}\cos\theta)v_i$$

since $\mathbf{v}_i$ and $\mathbf{v}_f$ have the same magnitude. Integrating with respect to φ from 0 to 2π causes the coefficients of $\mathbf{i}$ and $\mathbf{j}$ to vanish, and Eq. (8-48) becomes

$$d\mathbf{J} = dt\,nev_i\,\mathbf{k}\int_0^\pi(\cos\theta - 1)W_{if}(\theta)(mv_i)^2\sin\theta\,d\theta\,2\pi$$

This can be written as

$$d\mathbf{J} = -\mathbf{J}\,dt\frac{< 1 - \cos\theta >}{\tau_c} \qquad (8\text{-}49)$$

where the mean free time between collisions is defined by the relation

$$\frac{< 1 - \cos\theta >}{\tau_c} = \int(1 - \cos\theta)W_{if}(\theta)\,dV \qquad (8\text{-}50)$$

In Section 7-2, we showed that the relaxation time τ indicated the rate of decay of the current in accordance with

$$d\mathbf{J} = -\mathbf{J}\,dt/\tau \qquad (8\text{-}51)$$

Comparing this with Eq. (8-49) gives

$$\tau = \tau_c/< 1 - \cos\theta > \qquad (8\text{-}52)$$

PROBLEM 8-16 Under what circumstances does τ_c reduce to τ?

Consider now a semiconductor with N_A acceptors per unit volume. By making the simplification that the acceptors are uniformly distributed, the spacing between acceptors is

$$A^3 = 1/N_D$$

and each acceptor has an effective scattering-action radius of $A/2$.

PROBLEM 8-17 Show that the probability that an electron with an initial velocity $\mathbf{v}_i$ will be scattered is $8\, s\, ds/A^2$.

Hence
$$< 1 - \cos\theta > = 8\int_0^{A/2} (1 - \cos\theta) s\, ds$$

But
$$1 - \cos\theta = \frac{2}{1 + \tan^2\varphi}$$
$$= \frac{2}{1 + (s/k)^2}$$
$$= \frac{2}{1 + (4\pi\epsilon m v_i^2 s/e^2)^2}$$

so
$$< 1 - \cos\theta > = 2\left(\frac{E_A}{2E}\right)^2 \ln\left[1 + \left(\frac{2E}{E_A}\right)^2\right] \tag{8-53}$$

where
$$E = mv_i^2/2 \qquad E_A = 2e^2/4\pi\epsilon A$$

PROBLEM 8-18 Verify Eq. (8-53).

Then, by Eq. (8-52),

$$\tau_I = \frac{A}{2v}\left(\frac{2E}{E_I}\right)^2 \Big/ \ln\left[1 + \left(\frac{2E}{E_I}\right)^2\right]$$

where we write τ as τ_I and E_A as E_I to denote impurity scattering, and where v_i is written simply as v. This formula is due to Conwell and Weisskopf (1950),[11] who obtained it in the form

$$\tau_I = \frac{2A}{v\pi} \frac{(2E/E_I)^2}{\ln[1 + (2E/E_I)^2]} \tag{8-54}$$

through the use of somewhat different approximations.

PROBLEM 8-19 Verify that $\lambda = \frac{3}{2}$ in Eq. (7-117) for ionized-impurity scattering.

In order to calculate the impurity mobility μ_I from τ_I, we use the relations

$$\mu_I = \sigma_0/ne$$

and
$$\sigma_0 = 4g < \tau > \tag{8-55}$$

obtained by letting $B = 0$ in the conductivity tensor [Eq. (7-110)] and by assuming spherical energy surfaces. The integral involved in $< \tau >$ will have the logarithmic term in the integrand, but since this is a slowly-varying function of E, we will replace E in this term by its maximum value.

PROBLEM 8-20

(a) Show from Eq. (7-124) that $< \tau >$ has a term $E^3 e^{-E/kT}$ in the integrand.

(b) Show that this term has a maximum when

$$E = 3kT \tag{8-56}$$

When we use Eqs. (8-55) and (8-56), the Conwell-Weisskopf formula for impurity mobility becomes

$$\mu_I = \frac{2^{7/2}\epsilon^2(kT)^{3/2}}{\pi^{3/2}m^{1/2}e^3N_I} \frac{1}{\ln[1 + (3\epsilon kT/e^2N^{1/3})^2]} \tag{8-57}$$

PROBLEM 8-21 By using Eq. (7-131), verify Eq. (8-57) and find l.

The impurity mobility has also been calculated on the basis of the Born approximation by Brooks (1951)[12] and Herring (1954).[13] Their formula is

$$\mu_I = \frac{2^{7/2}}{\pi^{3/2}} \frac{\epsilon^2(kT)^{3/2}}{m_e{}^{1/2}e^3N_I} \frac{1}{\ln(1 + b) - b/(1 + b)} \tag{8-58}$$

where
$$b = \frac{6}{\pi} \frac{\epsilon m_e k^2 T^2}{n\hbar^2 e^2}$$

For $n = N_I$, the Brooks-Herring and Conwell-Weisskopf expressions give very similar results, as shown in Fig. 8-14, taken from Debye and Conwell (1954).[14] (Note that the two expressions are the same except for the logarithmic term, in any case.) The potential used in the Brooks-Herring formula is of the form

$$V \propto e^{-\alpha r}/r \tag{8-59}$$

and is called a *screened Coulomb potential* (α is the constant of Problem 1-21). The exponential term causes a much more rapid decrease of V with r than the ordinary Coulomb field, and Eq. (8-59) is used to account for screening effects. This concept is incorporated into Eq. (8-58), for when the conduction electron density is considerably less than the ionized impurity density, the resultant impurity mobility will be much lower. When $n = N_I$, the assumption made in deriving the Conwell-Weisskopf formula that each ion has a scattering radius roughly equal to one-half the impurity separation is valid, since the electrons screen the impurities. However, when $n \ll N_I$, the screening decreases, and the Brooks-Herring formula should be more correct.

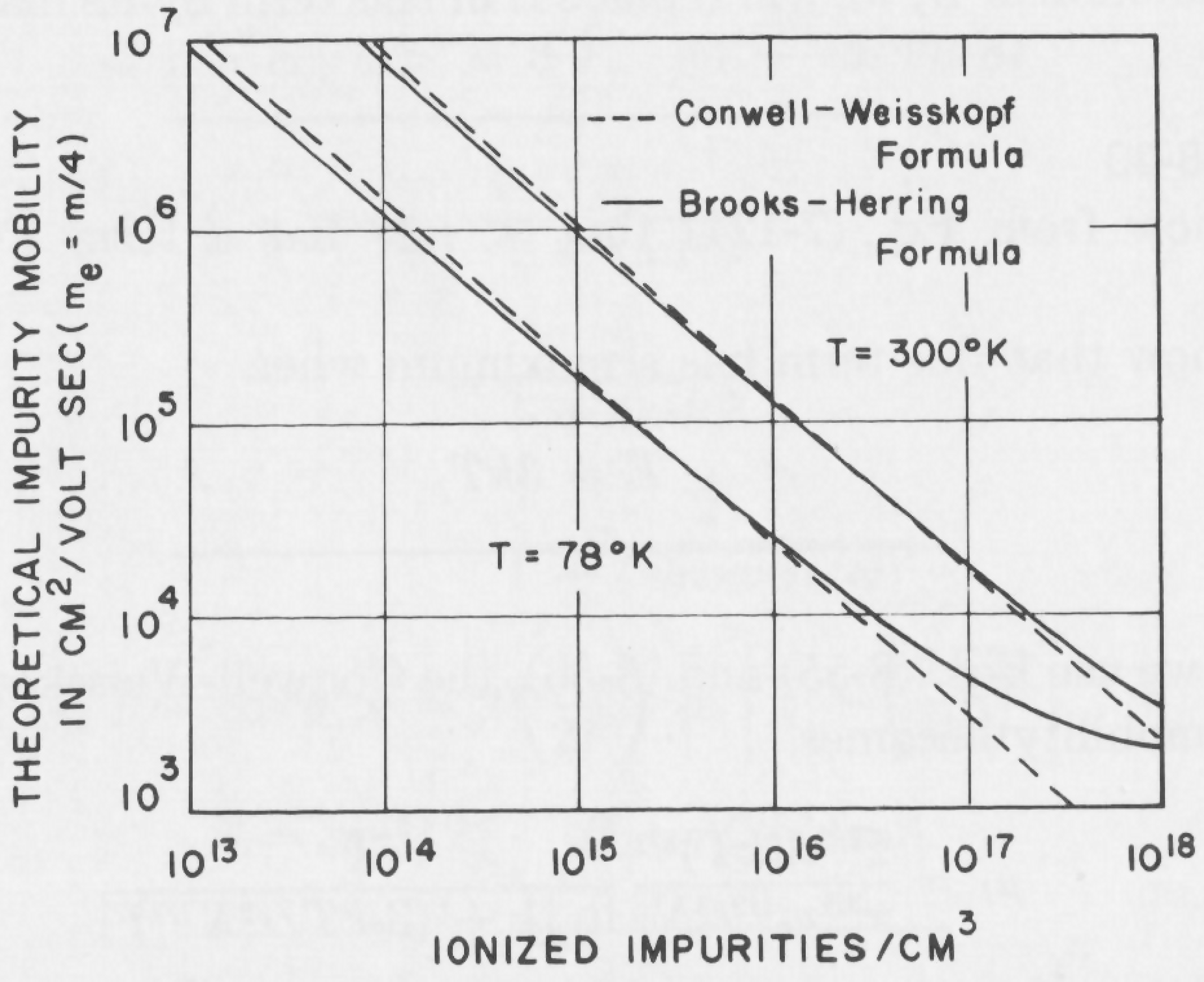


FIG. 8-14. Mobility as a function of impurity concentration. (From P. P. Debye and E. M. Conwell, Phys. Rev. **93**, 693, 1954, with permission.)

The variation of mobility with temperature as predicted by Eq. (8-57) or Eq. (8-58) has proven to be fairly difficult to verify experimentally. Figure 8-15 shows the variation of resistivity with temperature for a number of tellurium samples of various purities, as measured by Long and Kizilos (1960).[15] The curves labelled "as-grown" are for small-diameter crystals; the other samples were cleaved from larger crystals. Below a temperature of about 40°K, the low-resistivity samples (the most impure) can be assumed to have a constant free hole density. The variation of ρ with T should be exponential but instead, we see that the change is the much more gradual $T^{3/2}$ variation, as predicted.

We note that at temperatures above 40°K, the variation changes to $T^{-3/2}$, which indicates the onset of *lattice scattering*. That is, at higher temperatures, the vibrations of the atom in the lattice become significant and overshadow the impurity scattering. Collisions of electrons or holes with these vibrating atoms involve a transfer of energy. A quantum of energy associated with those vibrations is called a *phonon*; it is customary to refer to lattice or thermal scattering as *electron-phonon collisions*. Since the study of these processes is quite an involved subject, we shall merely give an intuitive justification of the $T^{-3/2}$ law.

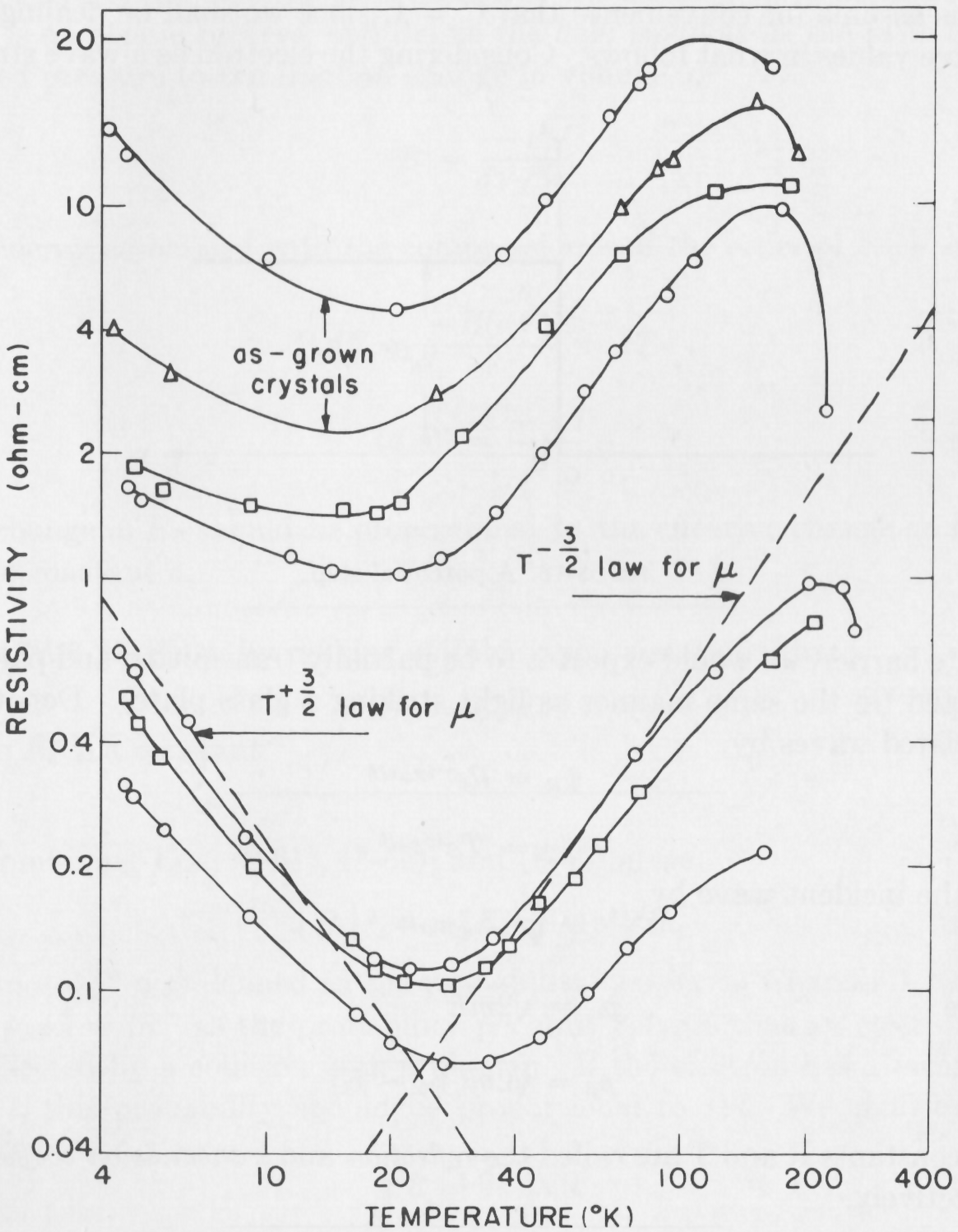


FIG. 8-15. Mobility as a function of temperature for tellurium crystals of various types. (From D. Long and B. Kizilos, Honeywell Research Center unpublished report, with permission.)

Again following Shockley (1950),[10] if the atoms in a crystal are vibrating, the width of the forbidden band is varying throughout the lattice. In order to solve the Schroedinger equation for such a situation, it is necessary to make the approximation that the change in band width is abrupt, rather than gradual, as shown in Fig. 8-16. The wave function representing the incident electron, from Eq. (1-71), is

$$\psi = Ce^{i\alpha x}$$

Let us assume for convenience that $C = 1$, since we shall be dealing with relative values in what follows. Considering the electron as a wave striking

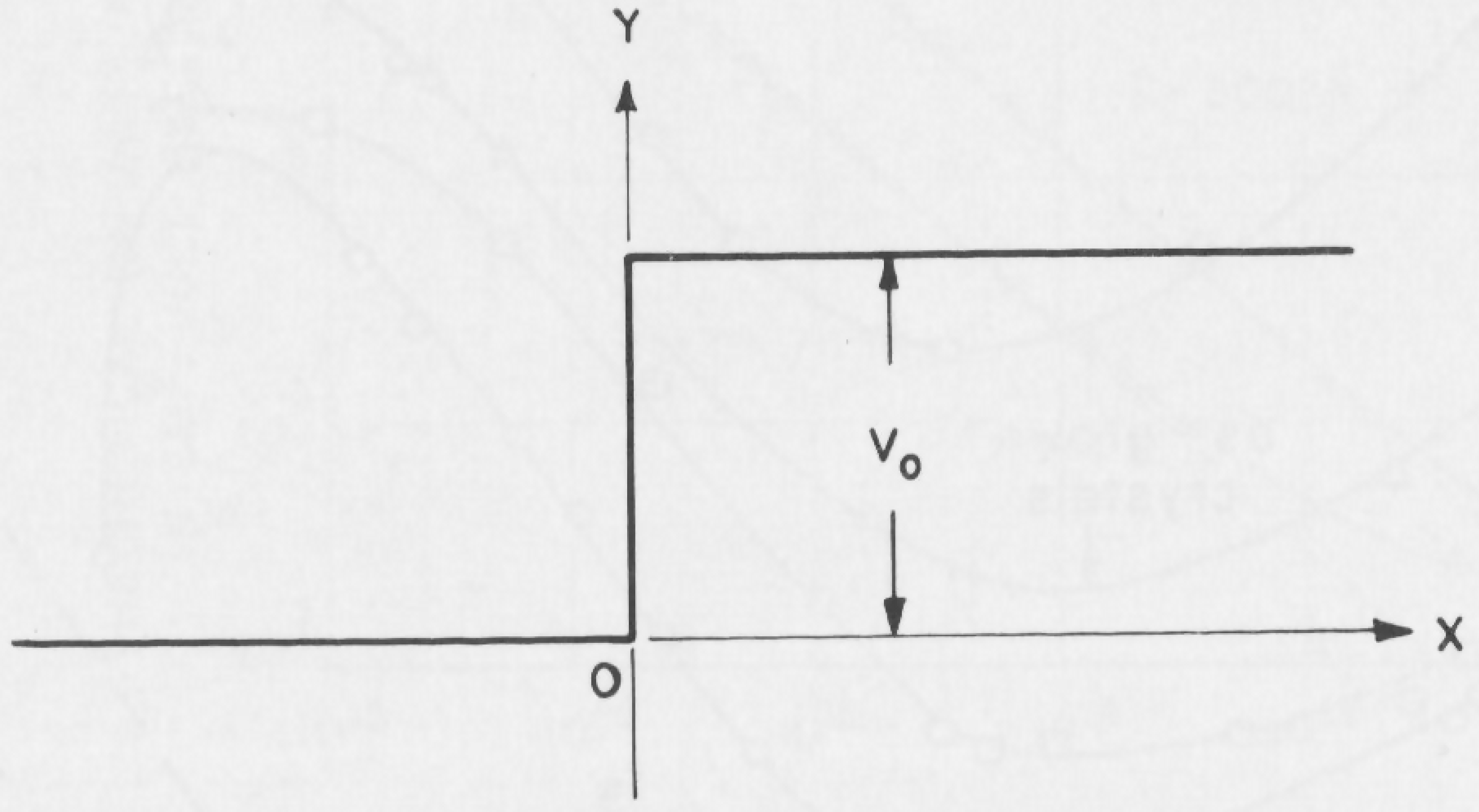


FIG. 8-16. A potential step.

a finite barrier, we would expect it to be partially transmitted and partially reflected (in the same manner as light striking a glass plate). Denote the associated waves by

$$\psi_R = Re^{-ip_\alpha x/\hbar}$$

$$\psi_T = Te^{ip_\beta x/\hbar}$$

and the incident wave by

$$\psi = e^{ip_\alpha l/\hbar}$$

where

$$p_\alpha = \sqrt{2mE}$$

$$p_\beta = \sqrt{2m(V_0 - E)}$$

The constants R and T are called the *reflection* and *transmission coefficients*, respectively.

PROBLEM 8-22 From the boundary conditions at $x = 0$, show that

$$|\,R\,|^2 = (p_\alpha - p_\beta)^2/(p_\alpha + p_\beta)^2 \tag{8-60}$$

If V_0 corresponds to a shift dE_C in energy for the bottom of the conduction band, then

$$dE_C = d(p^2/2m_e) = p\,dp/m_e \qquad ($$

and letting $dp = (p_\alpha - p_b)$ be small gives

$$|\,R\,|^2 = (dp)^2/4p^2 = m_e^2(dE_C)^2/4p^4 \tag{8-61}$$

where $p = p_\alpha \sim p_b$.

Let us assume that the deformation of the lattice by the atomic vibrations is an elastic process, and define the *bulk modulus* as the ratio of the applied pressure to the fraction change in volume, or

$$B = \frac{-P}{dV/V}$$

The energy associated with the compression is of the order of $P\,dV = kT$, and

$$P\,dV = \frac{-B(dV)^2}{V} = kT$$

or

$$(dV)^2 = \frac{-kT}{BV} \tag{8-62}$$

The change in E_C should be proportional to the effective change da in the lattice constant a.

PROBLEM 8-23 Show by making suitable approximations that

$$dE_C = E_a\,dV/V \tag{8-63}$$

where E_a is a constant.

Combining Eqs. (8-61), (8-62), and (8-63) gives

$$|\,R\,|^2 = m_e^2E_a^2kT/4p^2VB \tag{8-64}$$

Since $\psi\psi^*$ was defined as the probability-density in Chapter 1, we can take $\psi_R\psi_R^* = |R|^2$ as the probability per unit volume that an electron will be reflected by a collision with a phonon. If the electron has a mean free path l, this probability should be proportional to $1/l$. We shall express this as

$$|\,R\,|^2 = b\lambda/l \tag{8-65}$$

where b is a proportionality constant and $\lambda = p/h$. Combining Eqs. (8-64) and (8-65) gives

$$l = 4b^2h^4B/m_e^2\mathcal{E}_a^2kT \tag{8-66}$$

PROBLEM 8-24 Show that Eq. (8-66) leads to a temperature-variation of *lattice mobility* of the form

$$\mu_L = C/T^{3/2} \tag{8-67}$$

This variation is shown by some of the samples in Fig. 8-12; Fig. 8-17 shows the Hall mobility for three arsenic-doped germanium samples of Debye and Conwell (1954).[14] The top curve represents the purest sample,

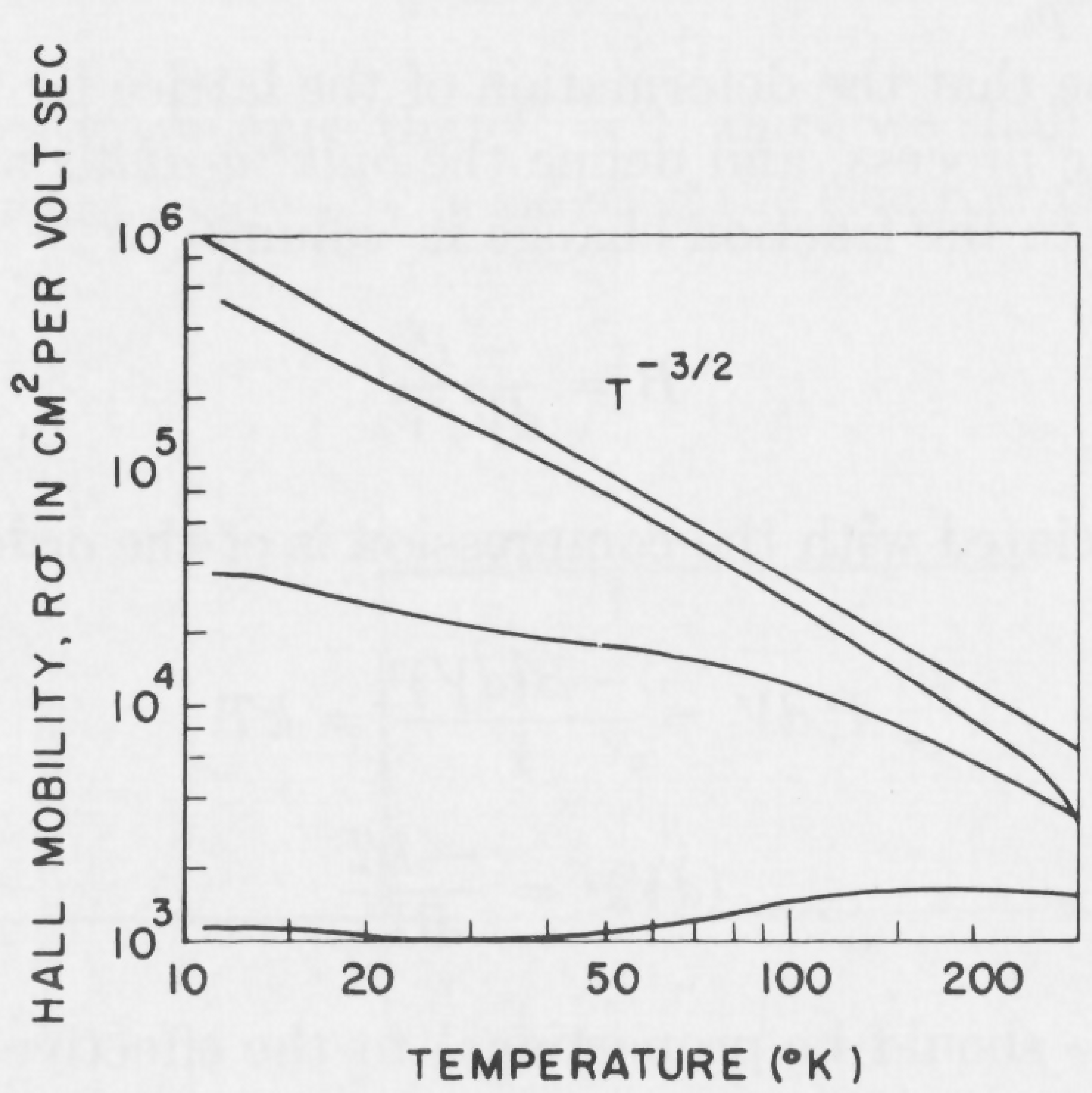


FIG. 8-17. Hall mobility as a function of temperature for N-type germanium. The purest sample is at the top. (From Debye and Conwell, *op. cit.*)

and the bottom one is for a sample which is degenerate over almost the entire temperature range.

The type of lattice scattering we have been considering is known as *accoustical* scattering, and we shall now show the source of this name. We can consider a vibrating lattice to a first approximation as a group of atoms executing harmonic motion and hence equivalent to a group of springs and masses. Figure 8-18a shows a one-dimensional model and, for the jth mass, from Section 1-5,

$$m\,\frac{d^2x_j}{dt^2} = -\kappa(x_j - x_{j-1}) - \kappa(x_j - x_{j+1})$$

where x denotes the displacement from equilibrium. This can be written

$$(m/\kappa)\,\frac{d^2x_j}{dt^2} = x_{j+1} - 2x_j + x_{j-1}$$

and we try a solution:

$$x = Ae^{i(qja-\omega t)}$$

PROBLEM 8-25 Show that the solution above leads to

$$m\omega^2/4\kappa = \sin^2 qa/2$$

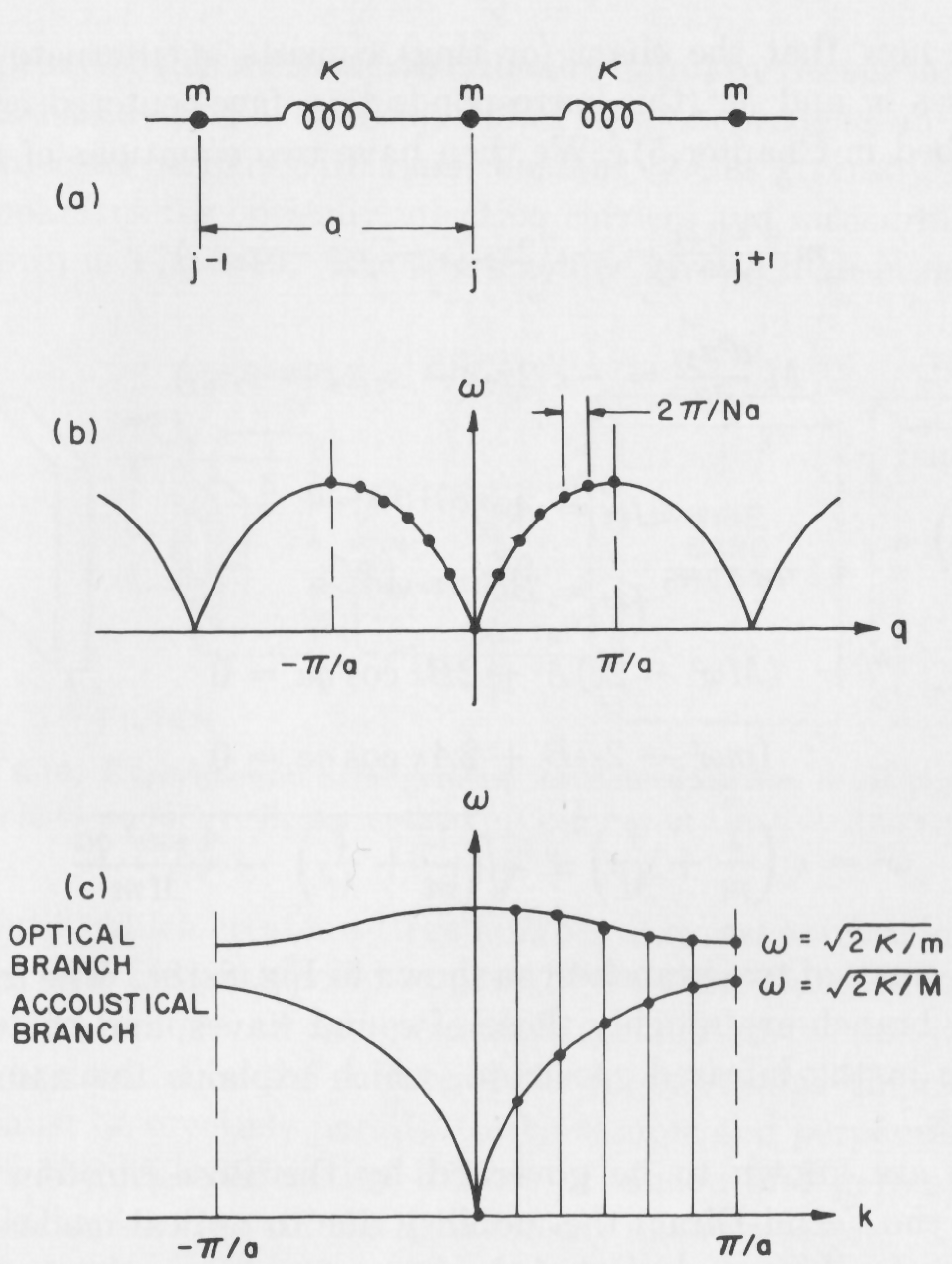


FIG. 8-18. The allowed frequencies of a vibrating chain.

Problem 8-25 shows that ω varies as

$$\omega = \omega_{\max} \sin qa/2$$

where

$$\omega_{\max} = 2\sqrt{\kappa/m}$$

If we bend the chain around into a ring of N atoms, as we did in Section 3-1, we once again have the cyclic boundary conditions.

$$x_{j+N} = x_j$$

and

$$e^{iqNa} = 1$$

or

$$q_n = 2\pi n/Na \qquad (n = 0, \pm 1, \ldots, \pm N)$$

Therefore, q, the phonon propagation constant, is quantized and so is ω, as shown in Fig. 8-18b, and we have phonon frequency bands which are analogous to the electron energy bands.

Suppose now that the chain (or ring) consists of alternate light and heavy masses m and M (this corresponds to a face-centered crystal like KCl described in Chapter 5). We then have two equations of motion:

$$m\frac{d^2x_{2j+1}}{dt^2} = -\kappa(2x_{2j+1} - x_{2j} - x_{2j+2})$$

$$M\frac{d^2x_{2j}}{dt^2} = -\kappa(2x_{2j} - x_{2j-1} - x_{2j+1})$$

with solutions:

$$x_{2j+1} = Ae^{i[(2j+1)qa-\omega t]}$$

$$x_{2j} = Be^{i(2jqa-\omega t)}$$

or

$$(M\omega^2 - 2\kappa)A + 2B\kappa\cos qa = 0$$

$$(m\omega^2 - 2\kappa)B + 2A\kappa\cos qa = 0$$

from which $$\omega^2 = \kappa\left(\frac{1}{m} + \frac{1}{M}\right) \pm \sqrt{\left(\frac{1}{m} + \frac{1}{M}\right)^2 - \frac{4\sin^2 qa}{Mm}}$$

This gives a curve of two branches, as shown in Fig. 8-18c. The frequencies of the lower branch are roughly those of sound waves, and the upper frequencies are in the infrared spectrum, which explains the names of the two branches.

Phonons are known to be governed by the Bose-Einstein statistics rather than the Fermi-Dirac; the mobility due to optical-mode scattering can be shown to obey a relation of the form

$$\mu_O = B(e^{\theta/T} - 1)$$

where θ is about 650°K.

The interpretation of experimental data in samples in which more than one kind of scattering is present is quite difficult. In such cases, it is found that μ may vary according to some power of T other than ± 1.5. Other types of scattering, such as those due to neutral impurities and to lattice imperfections, must also be considered. For a more comprehensive discussion, the review article of Brooks (1955)[16] is highly recommended.

8-6 Determination of Minority Carrier Lifetime

Our consideration of the important parameters governing the operation of semiconductor devices will be concluded with a few practical ideas on the problem of lifetime measurement. The suggestions to be given here are taken from an *I.R.E.* Standard prepared by Blakemore (1961).[18] The method recommended is the technique of *photoconductive decay*.

This theory of this method, described in detail by Stevenson and Keyes (1955),[17] is based on Eq. (4-106). This relation predicts an exponential decay of injected carriers with time constant $1/\nu$, as given by Eq. (4-112).

The apparatus for optically injecting carriers and measuring the decay time is shown in Fig. 8-19. The spark source gives a high-intensity, short-duration pulse, which creates a large number of excess carriers in the body of the sample. The filter is a polished piece of the same material as the sample and is used to remove all non-penetrating radiaton, which would cause carriers to be created on the surface of the sample rather than in the bulk. The filter must be precisely parallel to the sample and perpendicular to the beam, since the high refractive indices of silicon and germanium cause considerable displacement of the light for any misorientation. The masking on the sample eliminates edge-effects, and the sample contacts must be ohmic. The surface must be sandblasted or ground to obtain a high surface recombination velocity, so the surface lifetime can be computed from Problem 4-23

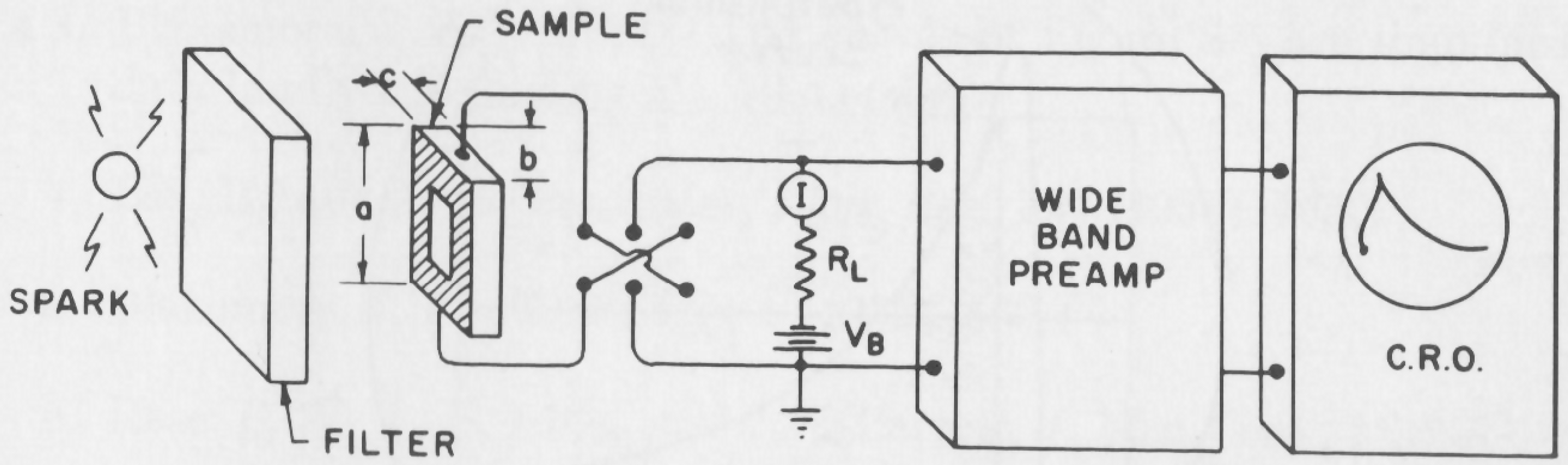


FIG. 8-19. Experimental arrangement for determination of lifetime by the photoconductive decay method. (Courtesy of Dr. J. S. Blakemore.)

$$\frac{1}{\tau_s} = \pi^2 D\left(\frac{1}{a^2} + \frac{1}{b^2} + \frac{1}{c^2}\right)$$

and the bulk lifetime τ can be found from the measured filament lifetime

$$\frac{1}{\tau} = \frac{1}{\tau_f} - \frac{1}{\tau_s} \tag{4-113}$$

The exponential decay of the excess carriers leads to an exponential decay of conductivity, and hence to the voltage across the sample for a constant current. This decay is matched against a standard exponential curve, which can be scribed on the face of the cathode ray oscilloscope. Figure 8-20a shows a typical scope face and exponential decay curve. When trapping (as discussed in Chapter 4) is present, the exponential character is altered, and the tail of Fig. 8-20b appears. Although the minority carrier decay and the process of emptying the traps are individually exponential, the sum of the two (or more) processes produces a non-exponential curve. Lifetime in samples for which the trapping is high cannot be determined by the photoconductive method. The theory of recombination and trapping is presented in great detail in two papers of Nomura and Blakemore (1958), (1961),[19,20] and some precautions on interpretation of the experimental data will be found in another paper by Blakemore and Nomura (1959).[21]

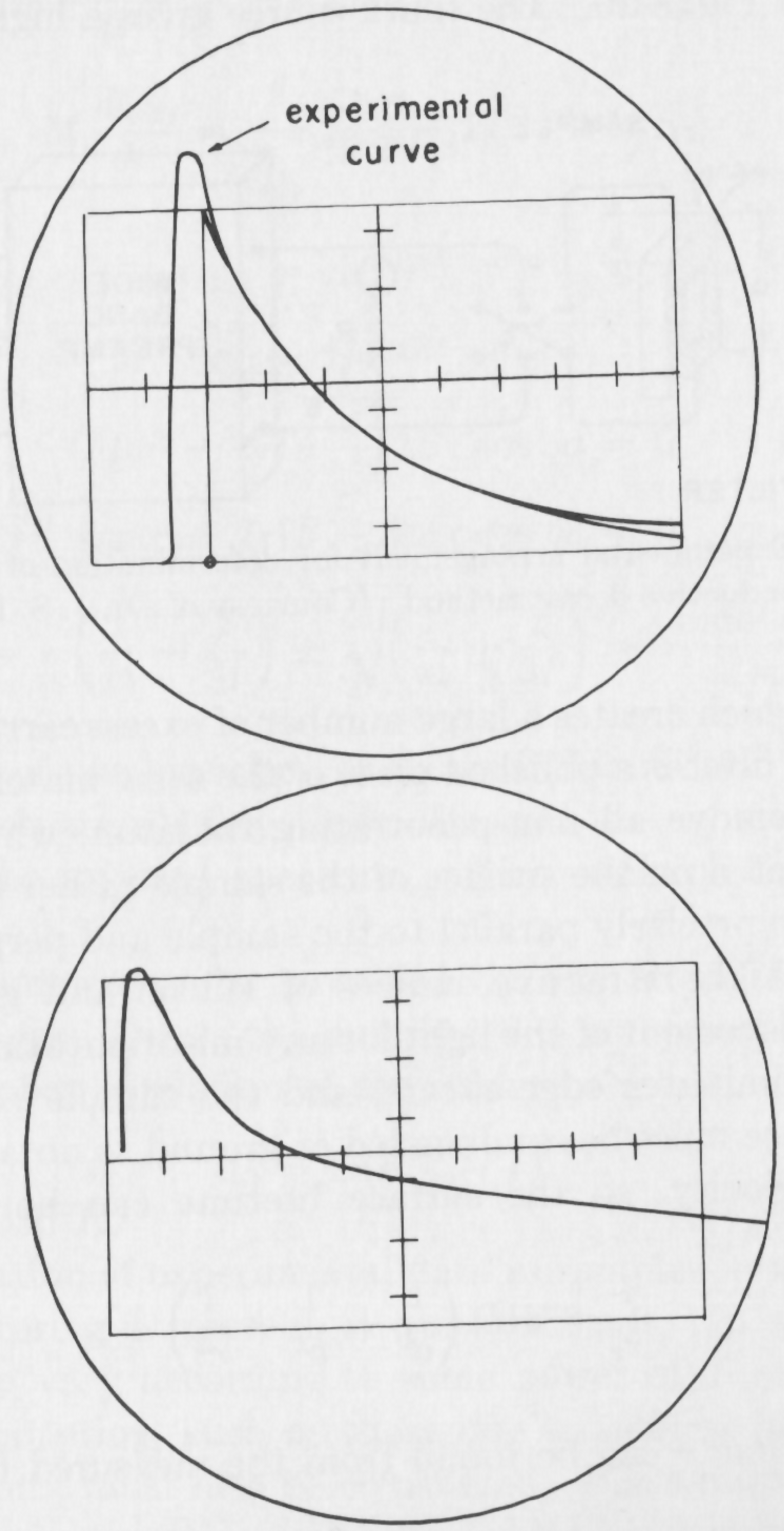


FIG. 8-20. Photoconductive decay curves (a) without trapping and (b) with trapping. (Courtesy of Dr. J. S. Blakemore.)

REFERENCES

1. Conwell, E. M., *Proc. I.R.E.*, **46** (1958), 1281.
2. Kohn, W., "Shallow Impurity States in Silicon and Germanium," *Solid State Physics*, Vol. 5, ed. F. Seitz and D. Turnbull.
3. Blakemore, J. S., (1961). The curves of Figure 8-5 are unpublished data kindly supplied by Dr. Blakemore.
4. Lax, B., and J. G. Mavroides, *Phys. Rev.*, **100** (1955), 1650.
5. Blakemore, J. S., *Phil. Mag.*, **4** (1959), 560.
6. Rose, F. W. G., E. Billig, and J. E. Parrott, *J. Elec. Cont.*, **3** (1957), 481.
7. Domenicali, C. A., *Phys. Rev.*, **112** (1958), 1863.
8. Joffe, A. V., *Physics of Semiconductors*. New York: Academic Press, 1960.
9. Born, M., *Atomic Physics*, 3rd ed. Glasgow: Blackie and Son, 1944.
10. Shockley, W., *Electrons and Holes in Semiconductors*. Princeton: Van Nostrand, 1950.
11. Conwell, E. and V. F. Weisskopf, *Phys. Rev.*, **77** (1950), 388.
12. Brooks, H., *Phys. Rev.*, **83** (1951), 879.
13. Herring, C., (1954), unpublished. (See reference 14.)
14. Debye, P. P. and E. M. Conwell, *Phys. Rev.*, **93** (1954), 693.
15. Long, D. and B. Kizilos, Honeywell Research Center Quarterly Report No. AF 49(638)-908, August 31, 1960.
16. Brooks, H., "Theory of the Electrical Properties of Germanium and Silicon," *Advances in Electronics and Electron Physics*, ed. L. Marton. New York: Academic Press, 1955.
17. Stevenson, D. T. and R. J. Keyes, *J. Appl. Phys.*, **26** (1955), 190.
18. Blakemore, J. S., *IRE Standards on Solid State Devices: Measurement of Minority-Carrier Lifetime in Germanium and Silicon by the Method of Photoconductive Decay*, IRE Standard 61 IRE 28.S3, 1961.

19. Nomura, K. C. and J. S. Blakemore, *Phys. Rev.*, **112** (1958), 1607.

20. Nomura, K. C. and J. S. Blakemore, *Phys. Rev.*, **121** (1961) 734.

21. Blakemore, J. S. and K. C. Nomura, *J. Appl. Phys.*, **31** (1959), 753.

GENERAL REFERENCES

Ziman, J. M., *Electrons and Phonons.* London: Oxford University Press, 1960. A thorough and rigorous treatment of the theory of scattering processes in semiconductors and metals.

APPENDIX

A-1 Taylor Series

Let us consider the problem of constructing a table of sines and cosines. It could be done by drawing a series of triangles, measuring the lengths of the sides, and calculating the ratios. This would be tedious and limited in accuracy, however, and what we would like is an analytic method. The most general algebraic formula we can have for sin x would be of the form

$$\sin x = a + bx + cx^2 + dx^3 + \cdots \qquad \text{(A-1)}$$

where $a, b, c, \ldots$ are constants to be determined. To find these constants, we put $x = 0$ in (A-1), obtaining $a = 0$. Next, differentiate (A-1)

$$\frac{d}{dx} \sin x = \cos x = b + 2cx + 3dx^2 + \cdots$$

and again let $x = 0$, which gives $b = 1$. Differentiating again,

$$\frac{d^2}{dx^2} \sin x = -\sin x = 2c + 6dx + \cdots$$

and at $x = 0$, we find that $c = 0$. If we continue this process, the result is

$$b = 1, \qquad d = -\tfrac{1}{6} = -\tfrac{1}{3!}, \qquad f = \tfrac{1}{120} = \tfrac{1}{5!}, \ldots$$

and

$$a = c = e = \cdots = 0$$

so

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots \tag{A-2}$$

Similarly,

$$\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \cdots \tag{A-3}$$

and

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots \tag{A-4}$$

The formulas given by Eqs. (A-2), (A-3), and (A-4) are called *Taylor expansions* or *series*. By substituting numerical values into these formulas, it will be seen that all terms after the first three or four are negligible; in using Taylor series, it is customary to retain only a few terms.

If we let $x = i\theta$ in Eq. (A-4), where $i = \sqrt{-1}$, then

$$e^{i\theta} = 1 + i\theta - \frac{\theta^2}{2!} - \frac{i\theta^3}{3!} + \cdots$$

and it follows from (A-2) and (A-3) that

$$e^{i\theta} = \cos\theta + i \sin\theta \tag{A-5}$$

Similarly, it can be shown that

$$e^{-i\theta} = \cos\theta - i \sin\theta \tag{A-6}$$

The validity of these various Taylor series and of the addition that led to (A-5) and (A-6) rests on the property of *convergence*. We shall not consider the rigorous definition of this term, but shall merely reassure ourselves that tables give finite values for sin x and cos x for any finite x.

The above procedure can be used to establish the form of a Taylor series for any function of x, which is then

$$f(x) = f(0) + \frac{x}{1!}\frac{df}{dx}(0) + \frac{x^2}{2!}\frac{d^2f}{dx^2}(0) + \cdots \tag{A-7}$$

and it is seen that Eqs. (A-2) and (A-3) are examples of (A-7).

If, instead of using $x = 0$, we let $x = a$, the result will be

$$f(x) = f(a) + \frac{(x-a)}{1!}\frac{df}{dx}(a) + \frac{(x-a)^2}{2!}\frac{d^2f}{dx^2}(a) + \cdots \tag{A-8}$$

For a function of two variables, the Taylor series is

$$\begin{aligned} f(x, y) = f(a, b) &+ (x-a)\frac{\partial f}{\partial x}(a, b) + (y-b)\frac{\partial f}{\partial y}(a, b) \\ &+ \frac{1}{2!}\left[(x-a)^2 \frac{\partial^2 f}{\partial x^2}(a, b) + 2(x-a)(y-b)\frac{\partial^2 f}{\partial x \partial y}(a, b)\right. \\ &\left. + (y-b)^2 \frac{\partial^2 f}{\partial y^2}(a, b)\right] + \cdots \end{aligned} \tag{A-9}$$

and the extension to three variables is obvious.

A-2 Vector Algebra

A *vector* is defined as a physical quantity possessing magnitude, sense, and direction. Examples of vectors are force, velocity, and electric field intensity (see Chapter 1 for a description of these quantities). A vector is symbolized by an arrow in the following fashion:

(1) The length of the arrow indicates the magnitude of the vector.
(2) The slope of the arrow with respect to a set of coordinate axes indicates the direction.
(3) The point of the arrow shows the sense.

Scalar quantities are those possessing magnitude and sense, but not direction. Examples are mass, temperature, and speed. A vector quantity — for example, force — is usually printed in boldface type, **F**, and its magnitude is then denoted by F. In taking notes, the student may wish to use the convention of placing an arrow over the letter, $\vec{F}$, as shown in Fig. A1. Two vectors, **F** and **F′**, are said to be equal if they have the same magnitude, direction, and sense, regardless of whether or not they coincide.

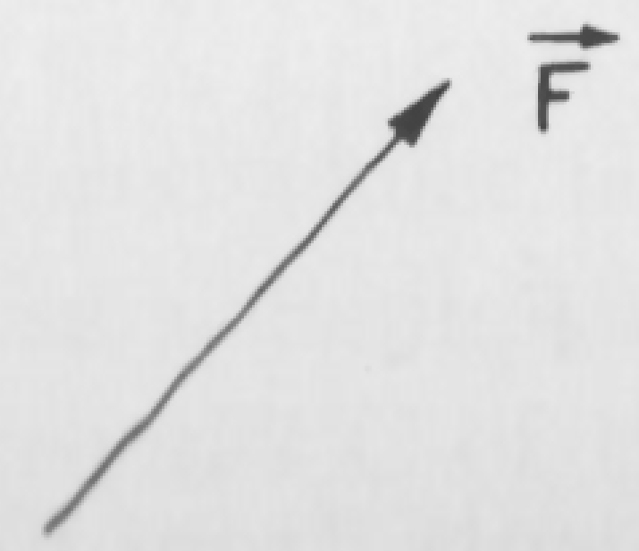


FIG. A1. A vector **F**.

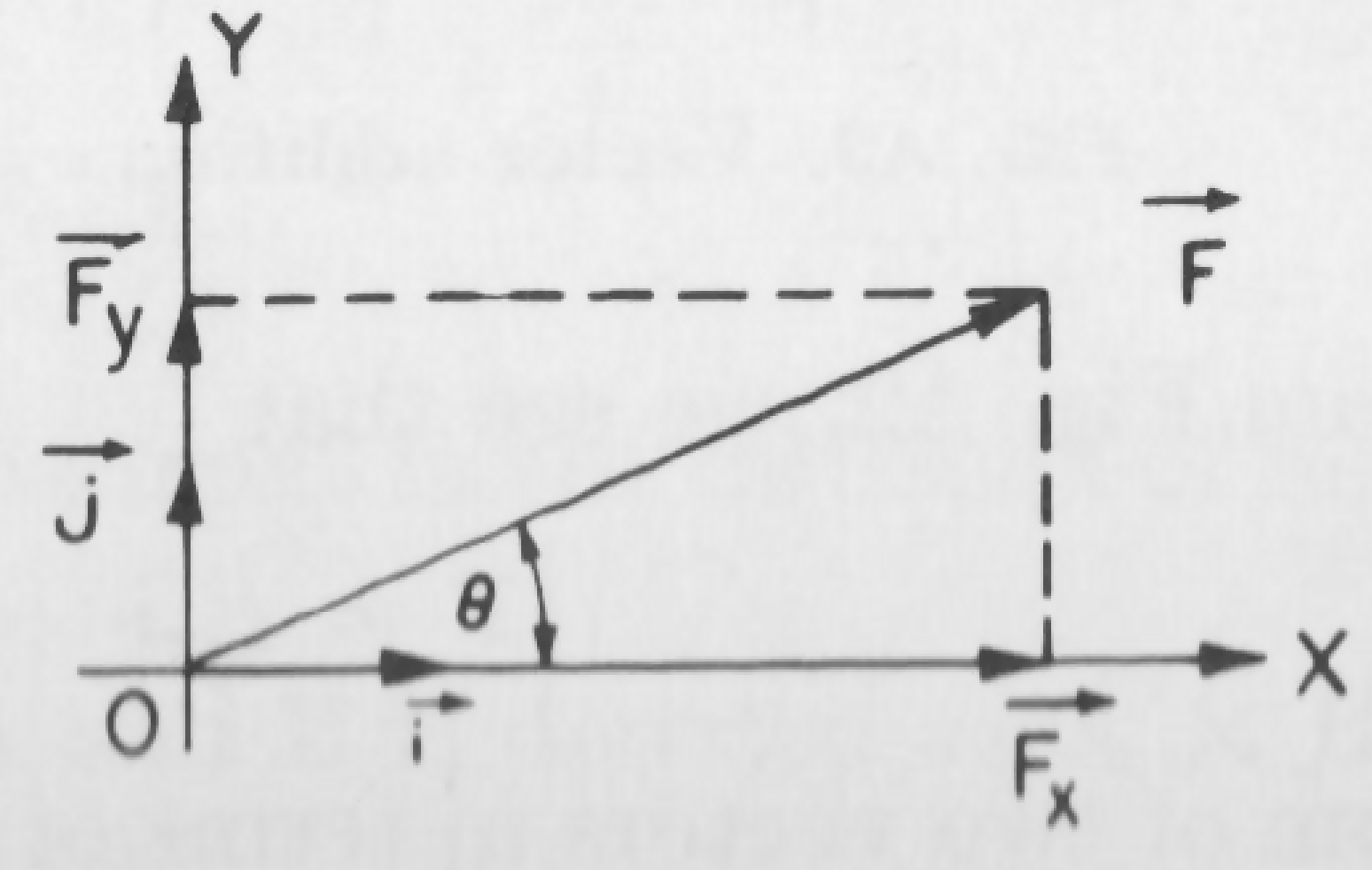


FIG. A2. The components of a vector.

The *components* of a vector are defined as its projections onto a set of coordinate axes. If a vector **F** has its initial point placed at the origin of the set of coordinates XOY (Fig. A2), the components of **F** are the vectors $\mathbf{F}_x$ and $\mathbf{F}_y$, where

$$F_x = F\cos\theta, \qquad F_y = F\sin\theta \tag{A-10}$$

Then

$$F^2 = F_x^2 + F_y^2, \qquad \theta = \arctan\,(F_y/F_x) \tag{A-11}$$

In three dimensions, the components of **F** with respect to a set of axes OXYZ are

$$F_x = F\cos\alpha, \qquad F_y = F\cos\beta, \qquad F_z = F\cos\gamma$$

where $\cos\alpha$, $\cos\beta$, and $\cos\gamma$ are the cosines of the angles which **F** makes with OX, OY, and OZ, respectively.

It is convenient at times to use *unit vectors.* Such a set of vectors lying along the three axes of the $OXYZ$ system are denoted **i**, **j**, and **k**, and their magnitudes are then $i = j = k = 1$. The unit vectors for the axes XOY are shown in Fig. A2.

The algebra of vectors is based on the following definitions:

(1) The vector $m\mathbf{F}$ is a vector whose direction and sense are the same as **F** and whose magnitude is mF.

(2) The vector $-\mathbf{F}$ has the same magnitude and direction as **F**, but the opposite sense.

(3) The sum **H** of two vectors **F** and **G** is obtained by placing the initial point of **G** at the terminal point of **F** (Fig. A3) and connecting the other two points. It is denoted by

$$\mathbf{F} + \mathbf{G} = \mathbf{H} \tag{A-12}$$

The difference of two vectors is then defined as the sum $\mathbf{F} + (-1)\mathbf{G}$.

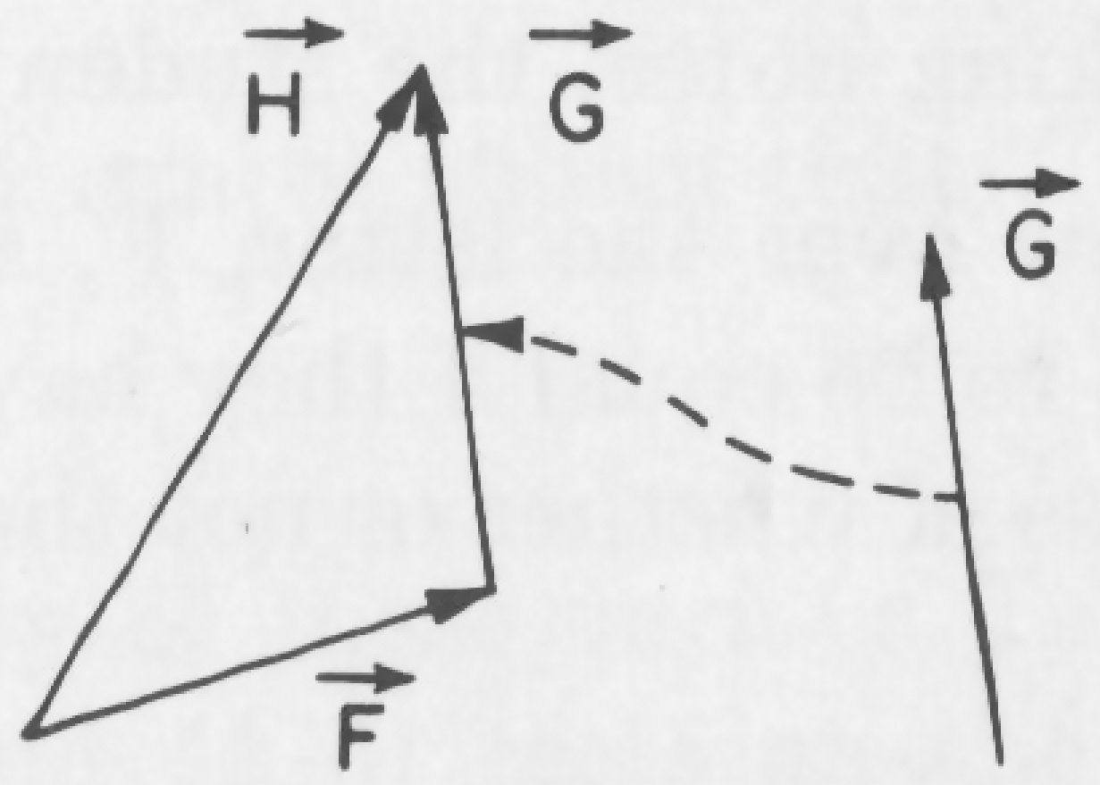


FIG. A3. Vector addition.

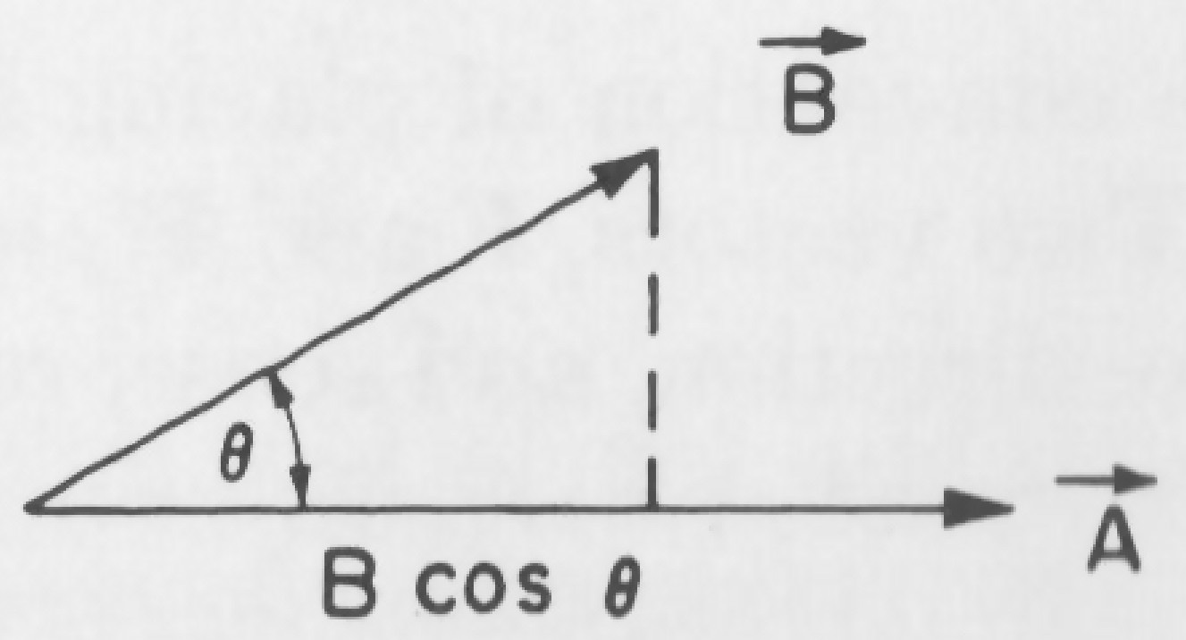


FIG. A4. The scalar product.

From Fig. A2, we see that

$$\mathbf{F} = F_x\mathbf{i} + F_y\mathbf{j} = \mathbf{F}_x + \mathbf{F}_y$$

The sum of two vectors in terms of their components, from Eqs. (A-11) and (A-12), is then

$$H^2 = (F_x + G_x)^2 + (F_y + G_y)^2 \tag{A-13}$$

The *scalar product* **C** of two vectors **A** and **B** is defined as the magnitude of **A** times the magnitude of the projection of **B** on **A** and is denoted by $\mathbf{A}\cdot\mathbf{B}$. From Fig. A4,

$$\mathbf{C} = \mathbf{A}\cdot\mathbf{B} = A(B\cos\theta) = AB\cos\theta = \mathbf{B}\cdot\mathbf{A} \tag{A-14}$$

The following special cases are important.

(1) **A** is parallel to **B**.
Then

$$\mathbf{A}\cdot\mathbf{B} = AB \tag{A-15}$$

(2) $\mathbf{A} = \mathbf{B}$.
Then

$$\mathbf{A}\cdot\mathbf{A} = A^2 \tag{A-16}$$

It is convenient to abbreviate $\mathbf{A}\cdot\mathbf{A}$ as $\mathbf{A}^2$, since the notation is consistent with the rest of vector algebra.

(3) **A** is perpendicular to **B**.
Then

$$\mathbf{A}\cdot\mathbf{B} = 0 \tag{A-17}$$

These relations show that

$$\mathbf{i}\cdot\mathbf{i} = \mathbf{j}\cdot\mathbf{j} = \mathbf{k}\cdot\mathbf{k} = 1$$

$$\mathbf{i}\cdot\mathbf{j} = \mathbf{j}\cdot\mathbf{k} = \mathbf{k}\cdot\mathbf{i} = 0$$

and from this, we have further

$$\begin{aligned}\mathbf{A}\cdot\mathbf{B} &= (A_x\mathbf{i} + A_y\mathbf{j} + A_z\mathbf{k})\cdot(B_x\mathbf{i} + B_y\mathbf{j} + B_z\mathbf{k})\\ &= A_xB_x + A_yB_y + A_zB_z \qquad \text{(A-18)}\end{aligned}$$

so

$$\mathbf{A}\cdot\mathbf{A} = A^2 = A_x^2 + A_y^2 + A_z^2 \tag{A-19}$$

as we would expect.

A convenient intuitive way to think of the scalar product is in terms of *work* W, which is defined by Eq. (1-5) as

$$W = \int \mathbf{F}\cdot d\mathbf{l}$$

where **F** is the force whose tangential component in the direction of motion, $F\cos\theta$, produces a displacement $d\mathbf{l}$ (Fig. A5).

The *vector product* **C** of two vectors **A** and **B** is denoted as $\mathbf{A}\times\mathbf{B}$, and is defined as the vector whose magnitude is

$$C = AB\sin\theta$$

where θ is the angle between A and B (Fig. A6), whose direction is perpendicular to the plane of **A** and **B**, and whose sense is such that **A**, **B**, and **C** form a *right-handed system*. By this, we mean that if **A** is rotated to **B**

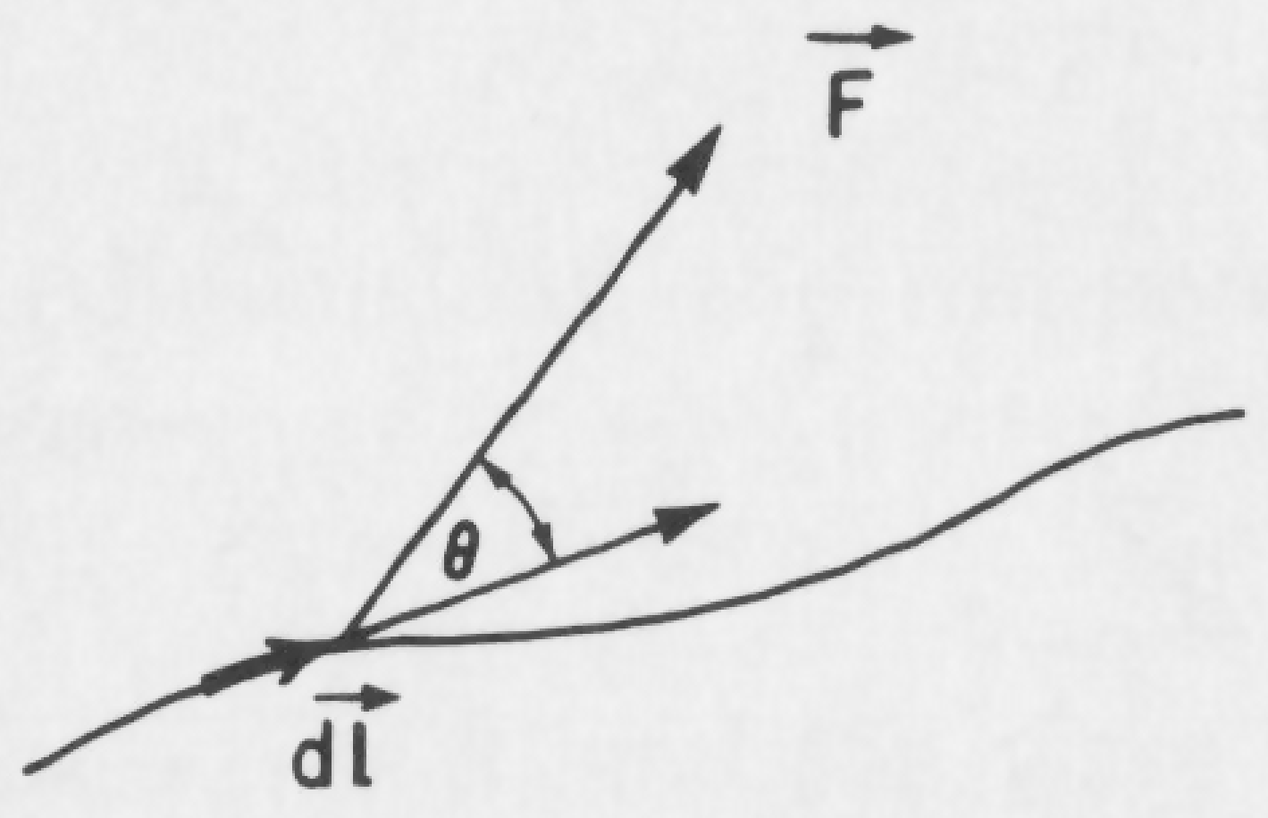


FIG. A5. Work represented by the scalar product.

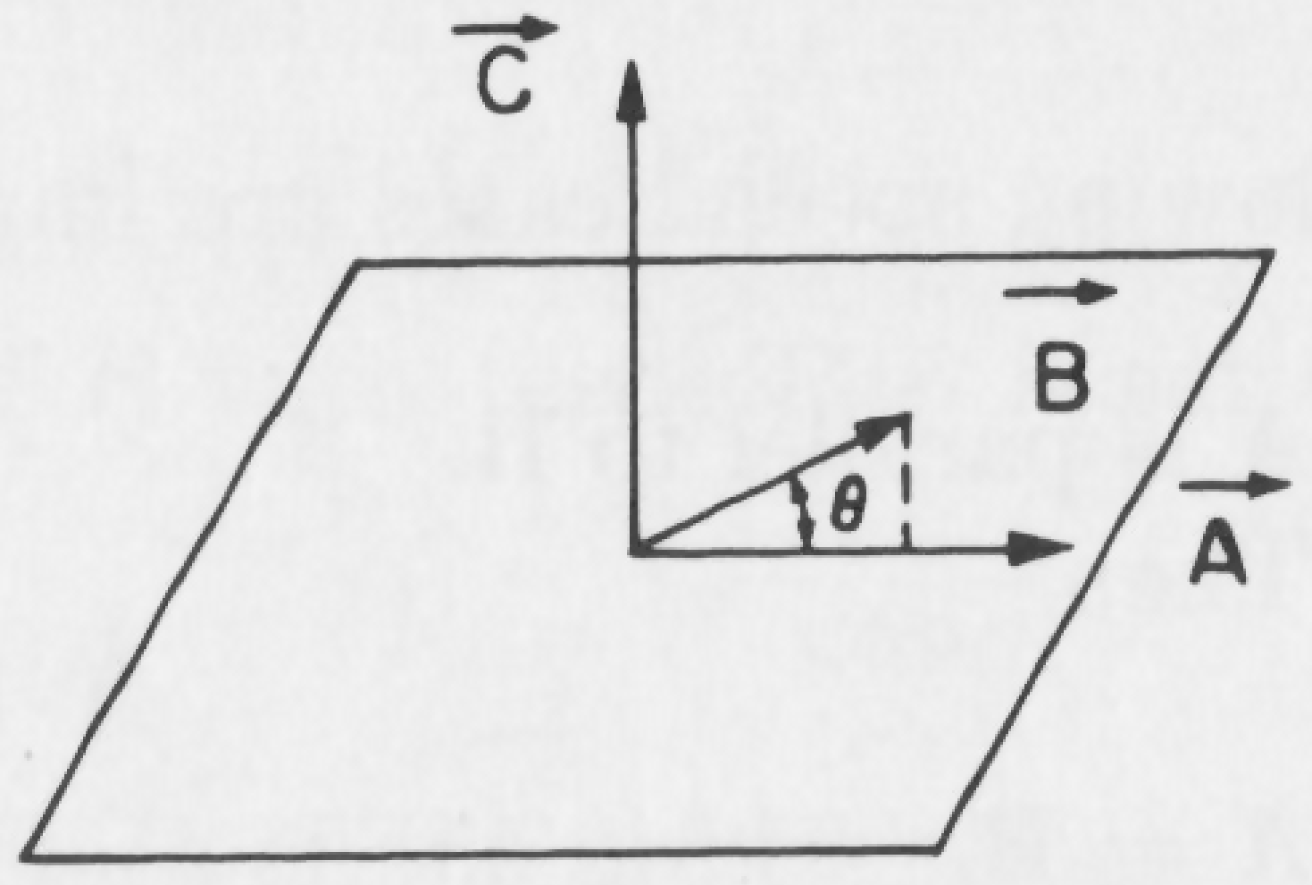


FIG. A6. The vector product.

about **C** as an axis, a right-handed screw would advance in the same sense as indicated by the arrow on **C**. It is customary to use right-handed sets of unit vectors, **i**, **j**, and **k**, in vector analysis.

The vector product can be interpreted geometrically as the area of the parallelogram determined by **A** and **B** in Fig. A6, since the altitude (shown dotted) is $B \sin \theta$. Hence, **C** can be regarded as the vector representation of this area.

From the definition of the vector product, we see that

$$\mathbf{A} \times \mathbf{B} = -\mathbf{B} \times \mathbf{A} \tag{A-20}$$

and from this, we have

$$\left.\begin{aligned} \mathbf{i} \times \mathbf{j} &= -\mathbf{j} \times \mathbf{i} = \mathbf{k} \\ \mathbf{j} \times \mathbf{k} &= -\mathbf{k} \times \mathbf{j} = \mathbf{i} \\ \mathbf{k} \times \mathbf{i} &= -\mathbf{i} \times \mathbf{k} = \mathbf{j} \end{aligned}\right\} \tag{A-21}$$

and

$$\mathbf{i} \times \mathbf{i} = \mathbf{j} \times \mathbf{j} = \mathbf{k} \times \mathbf{k} = 0 \tag{A-22}$$

Hence,

$$\begin{aligned} \mathbf{C} = \mathbf{A} \times \mathbf{B} &= (A_x\mathbf{i} + A_y\mathbf{j} + A_z\mathbf{k}) \times (B_x\mathbf{i} + B_y\mathbf{j} + B_z\mathbf{k}) \\ &= \mathbf{i}(A_yB_z - A_zB_y) + \mathbf{j}(A_zB_x - A_xB_z) + \mathbf{k}(A_xB_y - A_yB_x) \\ &= \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ A_x & A_y & A_z \\ B_x & B_y & B_z \end{vmatrix} \end{aligned} \tag{A-23}$$

where the vertical lines denote a determinant.

The vector product can be interpreted physically in terms of *torque*. A torque $\mathfrak{T}$ is a force which produces a rotation; it is defined as the product of the force and the perpendicular distance from the axis to the force-vector (Fig. A7), or

$$\mathfrak{T} = \mathbf{r} \times \mathbf{F}$$

The *scalar triple product* or *box product* of three vectors **A**, **B**, and **C** is defined as

$$\mathbf{C} \cdot (\mathbf{A} \times \mathbf{B}) = [\mathbf{ABC}]$$

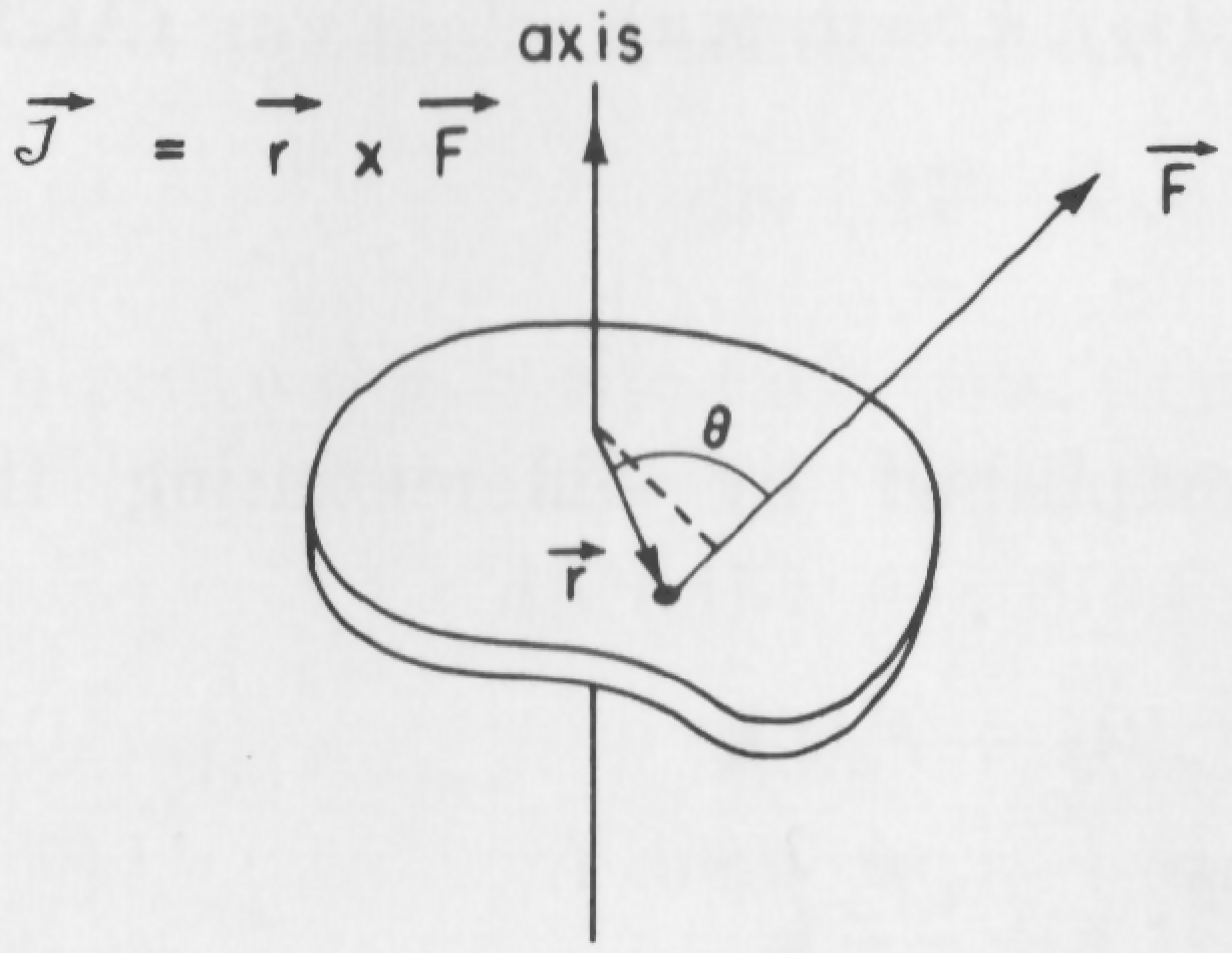


FIG. A7. Torque as a vector product.

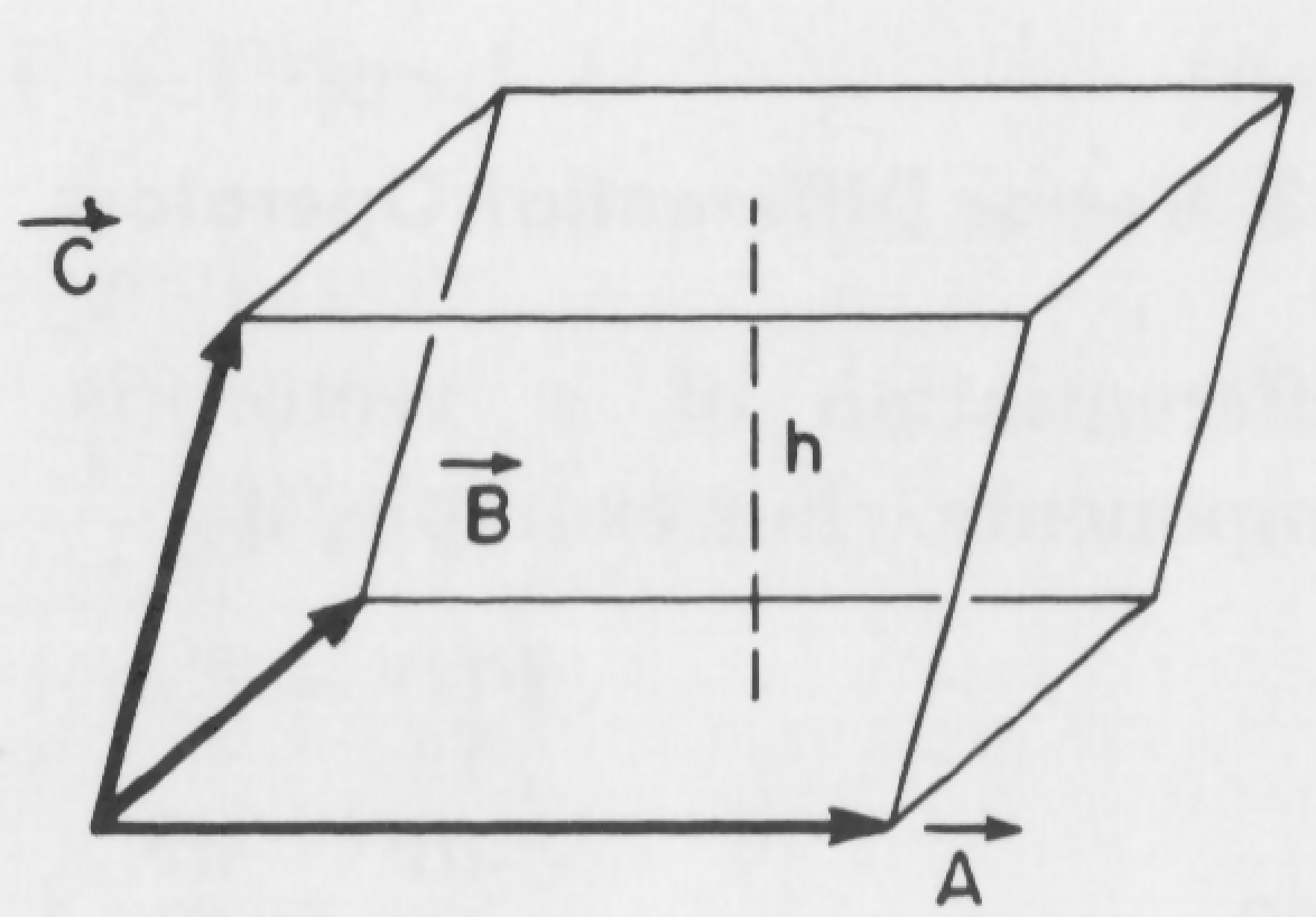


FIG. A8. The box product.

This can be interpreted as the volume of a parallelepiped (Fig. A8) which has three adjacent edges determined by **A**, **B**, and **C**. The area of the base equals the magnitude of $\mathbf{A} \times \mathbf{B}$, and the height is the projection of **C** on the normal to the base, so the volume is given by $\mathbf{C} \cdot \mathbf{A} \times \mathbf{B}$. We note that

$$[\mathbf{ABC}] = \begin{vmatrix} A_x & A_y & A_z \\ B_x & B_y & B_z \\ C_x & C_y & C_z \end{vmatrix} = [\mathbf{BCA}] = -[\mathbf{CBA}] = \text{etc.} \tag{A-24}$$

The expression [Eq. (A-18)] for the scalar product can be written in terms of the *Kronecker delta* δ_{ij}, defined by

$$\left.\begin{aligned} \delta_{ij} &= 1 \qquad \text{if } i = j \\ \delta_{ij} &= 0 \qquad \text{if } i \neq j \end{aligned}\right\} \tag{A-25}$$

Then Eq. (A-18) becomes

$$\mathbf{A} \cdot \mathbf{B} = \sum_{i,j} A_iB_j\delta_{ij} \qquad (i, j = x, y, z) \tag{A-26}$$

Similarly, the expression for the vector product [Eq. (A-22)] can be put in a simple form by the use of the *permutation symbols* ϵ_{ijk}, defined as follows:

$$\left.\begin{aligned} \epsilon_{ijk} &= 1 \qquad \text{when } ijk = xyz,\ yzx, \text{ or } zxy \\ \epsilon_{ijk} &= -1 \qquad \text{when } ijk = xzy,\ zyx, \text{ or } yxz \\ \epsilon_{ijk} &= 0 \qquad \text{otherwise} \end{aligned}\right\} \tag{A-27}$$

For example, $\epsilon_{xxy} = \epsilon_{xzz} = 0$

Then Eq. (A-23) can be written

$$C_i = \sum_{j,k} \epsilon_{ijk} A_j B_k \qquad (i, j, k = x, y, z) \tag{A-28}$$

A-3 Vector Differential Operators

Differentiation of a vector is accomplished by differentiating the components. For example, if

$$\mathbf{F}(t) = F_x(t)\mathbf{i} + F_y(t)\mathbf{j} + F_z(t)\mathbf{k}$$

then

$$\frac{d\mathbf{F}}{dt} = \frac{dF_x}{dt}\mathbf{i} + \frac{dF_y}{dt}\mathbf{j} + \frac{dF_z}{dt}\mathbf{k}$$

Consider now a vector whose magnitude, sense, and direction depend on its position in space. An example of such a vector is the electric field intensity $\boldsymbol{\mathcal{E}}$ existing near a group of charged bodies. The *divergence* of $\boldsymbol{\mathcal{E}}$ is defined and denoted as follows:

$$\operatorname{div} \boldsymbol{\mathcal{E}} = \frac{\partial \mathcal{E}_x}{\partial x} + \frac{\partial \mathcal{E}_y}{\partial y} + \frac{\partial \mathcal{E}_z}{\partial z} = \nabla \cdot \boldsymbol{\mathcal{E}} \tag{A-29}$$

where the symbol ∇ (sometimes read as *del*) is the vector differential operator

$$\nabla = \mathbf{i}\frac{\partial}{\partial x} + \mathbf{j}\frac{\partial}{\partial y} + \mathbf{k}\frac{\partial}{\partial z} \tag{A-30}$$

Now consider a scalar quantity which is a function of position, such as the potential energy $V(x, y, z)$ due to the gravitational attraction at various geographical locations. The *gradient* of V is defined as the vector

$$\mathbf{grad}\ V = \nabla V = \frac{\partial V}{\partial x}\mathbf{i} + \frac{\partial V}{\partial y}\mathbf{j} + \frac{\partial V}{\partial z}\mathbf{k} \tag{A-31}$$

These two operations can be combined to form the *Laplacian*, defined by

$$\operatorname{div} \mathbf{grad}\ V = \nabla \cdot \nabla V = \nabla^2 V = \frac{\partial^2 V}{\partial x^2} + \frac{\partial^2 V}{\partial y^2} + \frac{\partial^2 V}{\partial z^2} \tag{A-32}$$

The *curl* of a vector $\mathbf{A}$ is defined by

$$\operatorname{curl} \mathbf{A} = \nabla \times \mathbf{A} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \dfrac{\partial}{\partial x} & \dfrac{\partial}{\partial y} & \dfrac{\partial}{\partial z} \\ A_x & A_y & A_z \end{vmatrix}$$

Since this operator is encountered only briefly in this book during the course of a manipulation, it will not be discussed in detail.

Some useful relations exist among the vector operators. For example, if A is any scalar quantity and $\mathbf{F}$ is any vector, then

$$\operatorname{div}(A\mathbf{F}) = A \operatorname{div} \mathbf{F} + \mathbf{F} \cdot \mathbf{grad}\ A \tag{A-33}$$

To prove this, write each term in component form

$$\operatorname{div}(A\mathbf{F}) = \frac{\partial}{\partial x}(AF_x) + \frac{\partial}{\partial y}(AF_y) + \frac{\partial}{\partial z}(AF_z)$$

$$A \operatorname{div} \mathbf{F} = A\frac{\partial F_x}{\partial x} + A\frac{\partial F_y}{\partial y} + A\frac{\partial F_z}{\partial z}$$

$$\mathbf{F} \cdot \mathbf{grad}\ A = F_x\frac{\partial A}{\partial x} + F_y\frac{\partial A}{\partial y} + F_z\frac{\partial A}{\partial z}$$

Substituting these into Eq. (A-33) verifies the identity. Similarly, we can show that

$$\nabla(\mathbf{F} \cdot \mathbf{G}) = (\mathbf{F} \cdot \nabla)\mathbf{G} + \mathbf{F} \times (\nabla \times \mathbf{G}) + (\mathbf{G} \cdot \nabla)\mathbf{F} + \mathbf{G} \times (\nabla \times \mathbf{F}) \tag{A-34}$$

We should now like to give an intuitive, physical interpretation of the vector operators, starting with the gradient. In Figure A9 we have sketched

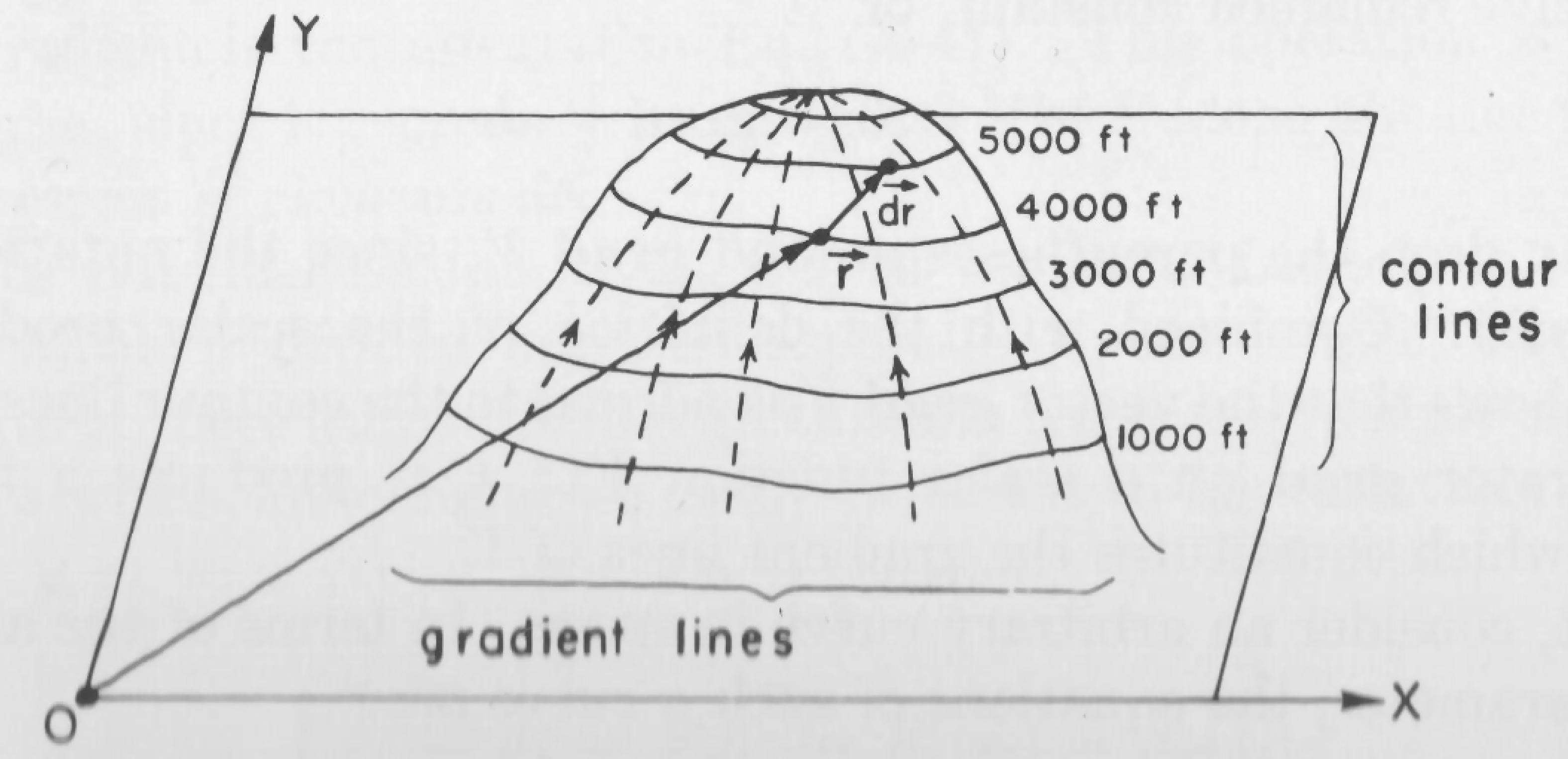


FIG. A9. Interpretation of the gradient.

a low hill; drawn on this hill are solid *contour lines*, with their height above sea level indicated. The dashed lines on the hill are a set of lines drawn to intersect the contour lines at right angles; they are called the *gradient lines*. The gradient lines represent the shortest paths from any level to the one immediately adjacent, and the contour lines are lines of constant gravitational potential energy V. Now the work done in getting to a specific point on the hill depends upon the location (x, y, z) of that point, which we indicate by writing V as $V(x, y, z)$ or $V(\mathbf{r})$, where $\mathbf{r}$ is the position-vector of this point. When a displacment $d\mathbf{r}$ is made from one contour line to another in an arbitrary direction, as shown in the figure, the corresponding change in V is

$$dV = V(\mathbf{r} + d\mathbf{r}) - V(\mathbf{r}) \tag{A-35}$$

We wish to expand the first term on the right by the use of Taylor's theorem. This can be done by expressing the vectors in component form.

$$\begin{aligned} V(\mathbf{r} + d\mathbf{r}) &= V(x + dx, y + dy, z + dz) \\ &= V(x, y, z) + \frac{\partial V}{\partial x}\,dx + \frac{\partial V}{\partial y}\,dy + \frac{\partial V}{\partial z}\,dz + \cdots \end{aligned}$$

If we neglect higher-order terms, Eq. (A-35) becomes

$$dV = \frac{\partial V}{\partial x}\,dx + \frac{\partial V}{\partial y}\,dy + \frac{\partial V}{\partial z}\,dz$$

and this is seen to be identical to

$$\begin{aligned} dV &= \left(\mathbf{i}\,\frac{\partial V}{\partial x} + \mathbf{j}\,\frac{\partial V}{\partial y} + \mathbf{k}\,\frac{\partial V}{\partial z}\right)\cdot(\mathbf{i}\,dx + \mathbf{j}\,dy + \mathbf{k}\,dz) \\ &= (\mathbf{grad}\ V)\cdot d\mathbf{r} \end{aligned} \tag{A-36}$$

If the displacement $d\mathbf{r}$ had been tangent to a contour line, the value of V would have remained constant, or

$$dV = 0 = \mathbf{grad}\ V\cdot d\mathbf{r} \tag{A-37}$$

(We may drop the parentheses around **grad** V, since the notation is unambiguous). Combined with the definition of the scalar product, Eq. (A-37) shows that the vector **grad** V is normal to the contour lines. Hence, the operator **grad** on a scalar function $V(x, y, z)$ produces a family of vectors which constitutes the gradient lines of V.

Next, consider an arbitrary curve in space. In terms of the arc-length s as a parameter, the equations of such a curve are

$$x = x(s), \qquad y = y(s), \qquad z = z(s)$$

The change in V corresponding to a change ds in s is

$$dV = \left(\frac{\partial V}{\partial x}\frac{dx}{ds} + \frac{\partial V}{\partial y}\frac{dy}{ds} + \frac{\partial V}{\partial z}\frac{dz}{ds}\right) ds \tag{A-38}$$

Now the derivatives dx/ds, dy/ds, and dz/ds are direction-cosines of the tangent to the curve, for they represent the limit of ratios of the form $|\Delta\mathbf{x}|/|\Delta\mathbf{s}|$, where $\Delta\mathbf{s}$ is a small vector along the curve and $\Delta\mathbf{x}$ is its x-component. If these direction-cosines are denoted by α, β, γ, then Eq. (A-38) becomes

$$dV = [(\partial V/\partial x)\alpha + (\partial V/\partial y)\beta + (\partial V/\partial z)\gamma]\,ds$$

Now let $d\mathbf{t}$ denote a small displacement tangential to the curve. Then

$$d\mathbf{t} = \mathbf{i}\alpha\,ds + \mathbf{j}\beta\,ds + \mathbf{k}\,\gamma ds$$

and

$$dV = (\mathbf{grad}\ V)\cdot d\mathbf{t}$$

Suppose the curve is now required to lie along one of the gradient lines in the figure. Then $d\mathbf{t}$ may be denoted by $d\mathbf{n}$, a displacement normal to a contour line. Further, **grad** V lies along $d\mathbf{n}$, as we showed above, so

$$dV = |\mathbf{grad}\ V|\,dn \tag{A-39}$$

That is, the scalar product reduces to the product of the magnitudes, since the vectors **grad** V and $d\mathbf{n}$ are colinear.

Finally, if we define force $\mathbf{F}$ as

$$\mathbf{F} = -\mathbf{grad}\ V \tag{A-40}$$

then, by Eq. (A-36),

$$V = -\int \mathbf{F}\cdot d\mathbf{r} \tag{A-41}$$

and this is in agreement with Eq. (1-5) of Section 1-3. Hence, the inverse of the gradient is the integral in Eq. (A-41). This operation is called a *line integral*, since it represents an integration of $\mathbf{F}$ along the line made up of a succession of elements $d\mathbf{r}$.

Let us now discuss the divergence in the same fashion. Consider a volume which is divided into strips parallel to the x-axis. Figure A10 shows one of these strips. Although this one has a very simple shape, the argument which follows can be easily extended to any arbitrary volume. Let $\mathbf{F}(x, y, z)$ be a vector function of position. By definition,

$$\int_{x_1}^{x_2} \frac{\partial F_x}{\partial x}\,dx = F_x\Big|_{x_1}^{x_2} = F_x(x_2) - F_x(x_1) \tag{A-42}$$

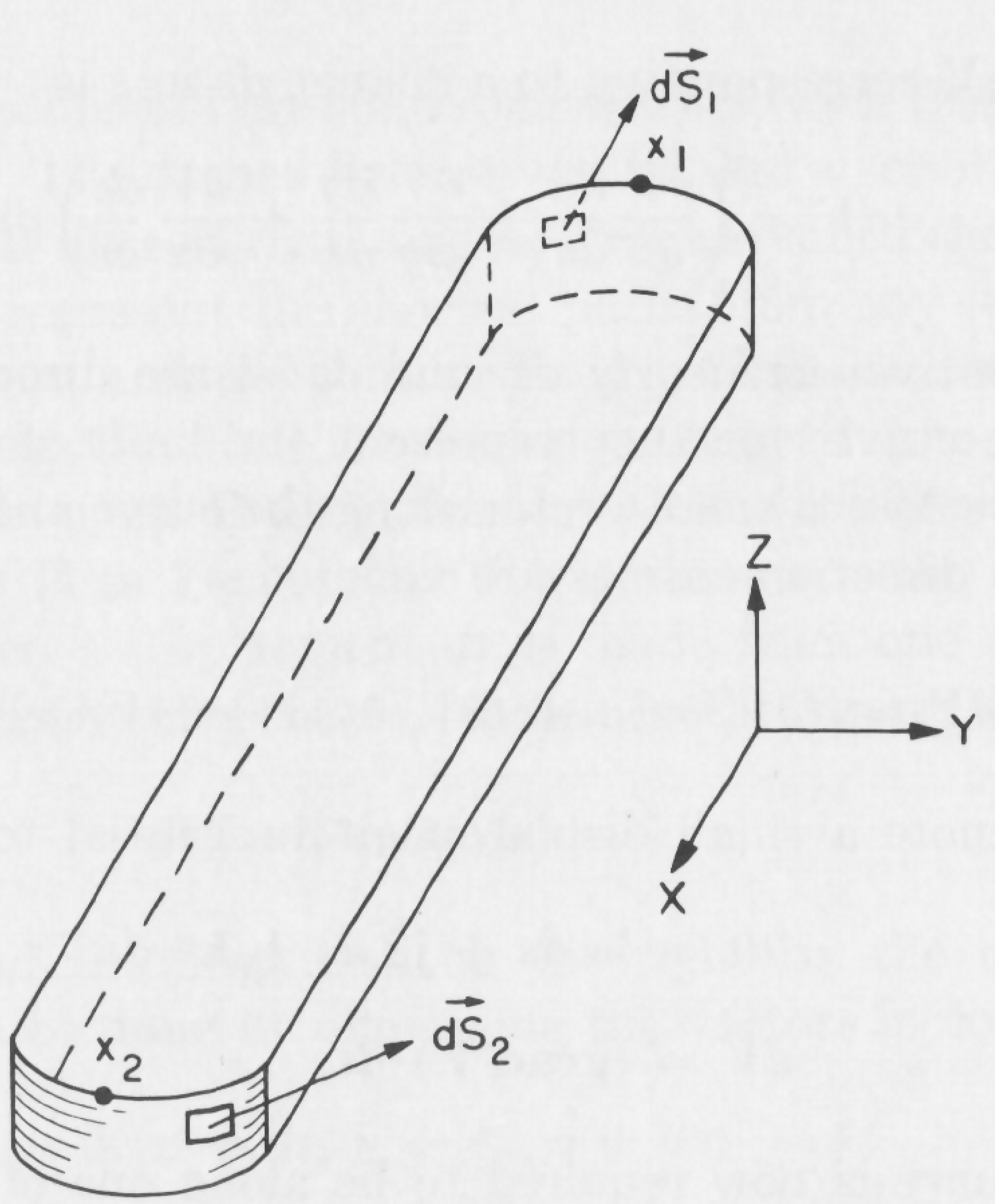


FIG. A10. Interpretation of the divergence.

If we regard the surface elements $d\mathbf{S}_1$ and $d\mathbf{S}_2$ as vectors, they can be expressed

$$dy\,dz = -\mathbf{i}\cdot d\mathbf{S}_1 = \mathbf{i}\cdot d\mathbf{S}_2 \tag{A-43}$$

since the projection of these elements on the x-axis is an area in the YOZ plane. Multiplying Eq. (A-42) by $dy\,dz$ and using Eq. (A-43) gives

$$\int \frac{\partial F_x}{\partial x}\,dx\,dy\,dz = \int F_{x2}\mathbf{i}\cdot d\mathbf{S}_2 + \int F_{x1}\mathbf{i}\cdot d\mathbf{S}_1 \tag{A-44}$$

The left-hand side of this equation is a *volume integral* and may be written as

$$\int \frac{\partial F_x}{\partial x}\,dx\,dy\,dz = \int_V \frac{\partial F_x}{\partial x}\,dV$$

The right-hand side is the sum of two *surface integrals;* it represents the integration of F_x over the surface bounding the volume V, and hence Eq. (A-44) becomes

$$\int_V \frac{\partial F_x}{\partial x}\,dV = \oint_S F_x\mathbf{i}\cdot d\mathbf{S} \tag{A-45}$$

where the symbol $\oint_S$ denotes integration over the entire closed surface S bounding the volume V. Adding, similarly, equations for F_y and F_z gives

$$\int_V \operatorname{div}\mathbf{F}\,dV = \oint_S \mathbf{F}\cdot d\mathbf{S} \tag{A-46}$$

This relation is known as the *divergence theorem,* and is the basis of our physical interpretation of the divergence.

Consider a point charge q and a volume V located near q (Fig. A11). This situation is like that of Fig. 4-2, but there the charge is inside the closed surface. By Eq. (A-46), we have

$$\int_V \operatorname{div}\boldsymbol{\mathcal{E}}\,dV = \oint_S \boldsymbol{\mathcal{E}}\cdot d\mathbf{S} \tag{A-47}$$

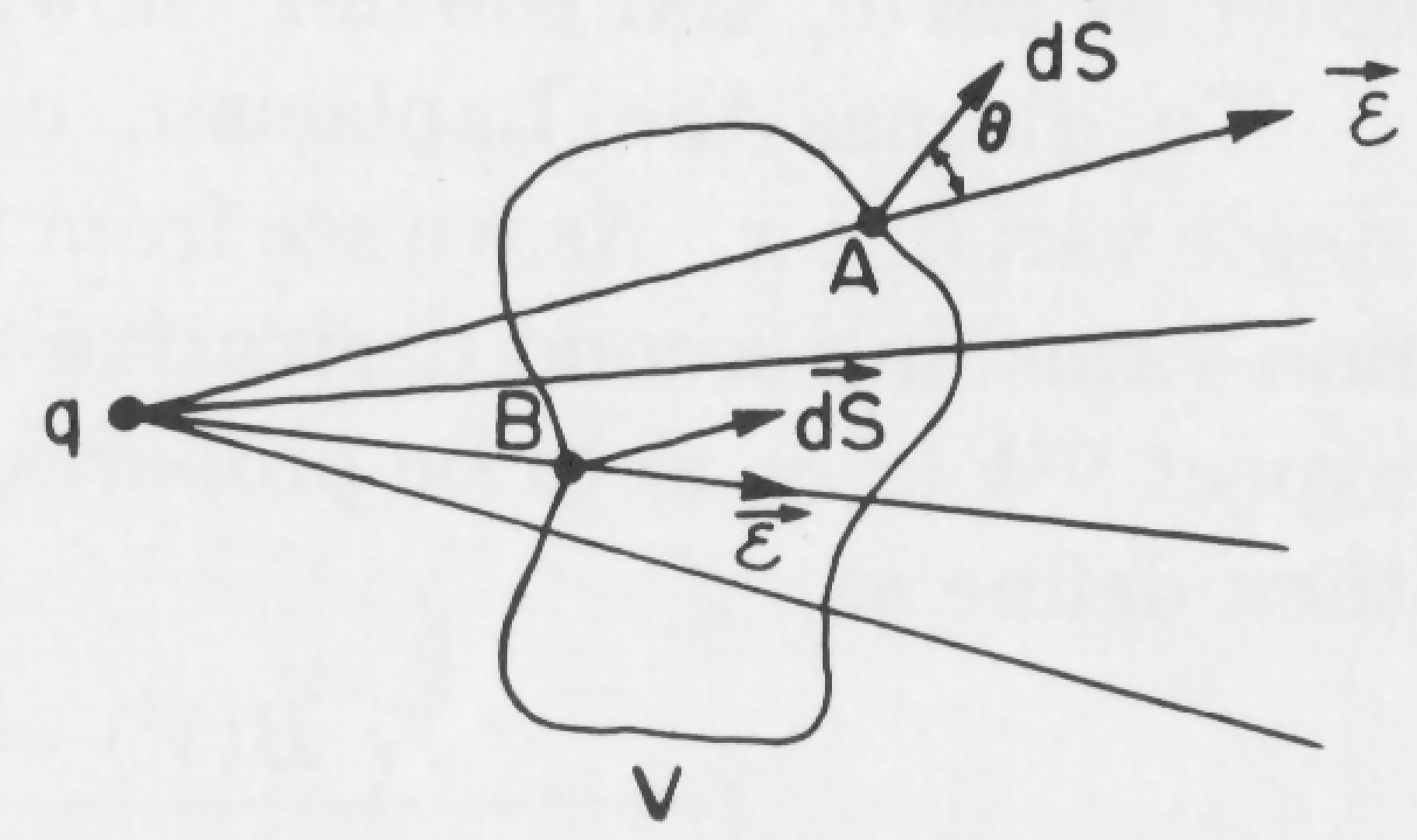


FIG. A11. The relation between the divergence and flux.

The surface integral represents the sum of all terms $\mathcal{E}\,dS\cos\theta$ for the entire surface; we see that at point A the vector $\boldsymbol{\mathcal{E}}$ points outwards, whereas at B it points inwards, so the contributions of these two points to the integral oppose one another. On the other hand, in Fig. 4-2, all the $\boldsymbol{\mathcal{E}}$ vectors point out, and the integral represents the sum of all the contributions. Combining Eqs. (A-47) and (4-4), we have

$$q/\epsilon = \int \operatorname{div}\boldsymbol{\mathcal{E}}\,dV = \oint \boldsymbol{\mathcal{E}}\cdot d\mathbf{S} \tag{A-48}$$

Gauss' law was derived for the situation where the charge is inside V; when it is outside, the contributions to the surface integral cancel in pairs, which agrees with the substitution of $q = 0$ in Eq. (A-48). Hence, we can think of the divergence of the field $\boldsymbol{\mathcal{E}}$ due to a charge q as the sum of all the normal components of $\boldsymbol{\mathcal{E}}$ across a surface S. This can also be thought of in terms of lines of force: when the surface surrounds the charge, the integral of the normal component of $\boldsymbol{\mathcal{E}}$ over the entire surface can be defined as the number of lines of force, or the *electric flux.* Then the divergence represents the "flow" of these lines of force from the charge q to the

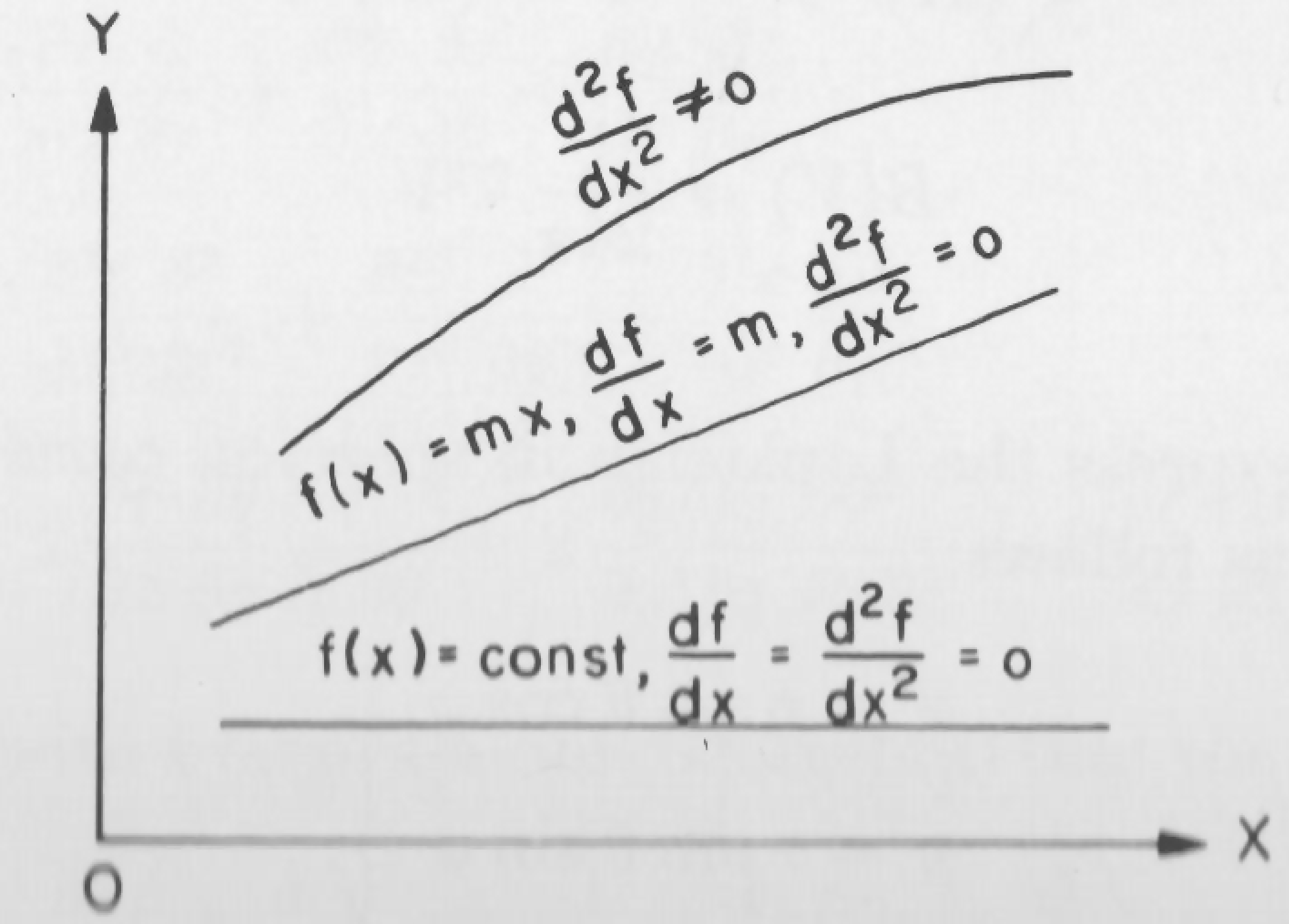


FIG. A12. The interpretation of the Laplacian.

surface where they are counted. If q is outside the surface, as many lines enter as leave, and the net "flow," or divergence, vanishes.

To discuss the Laplacian, consider first a function $y = f(x)$ of the single variable x. As we see from the curves in Fig. A12, this function has a non-vanishing second derivative only when it shows bulges, and we would expect $\nabla^2 V(x, y, z)$ to be proportional to the *bulginess* $B(V)$ of V, which we then define as

$$B(V) = \frac{V - < V >_S}{S} \tag{A-49}$$

where S is the area of a small spherical surface surrounding the point at which the bulginess is being measured, and $< V >_S$ denotes the average value of V over S. By a Taylor expansion, this becomes

$$B(V) = [V(\mathbf{r}) - < V(\mathbf{r}) + d\mathbf{r}\cdot\nabla V + \tfrac{1}{2}(d\mathbf{r}\cdot\nabla)^2 V + \cdots >_S]/4\pi(d\mathbf{r})^2 \tag{A-50}$$

where $d\mathbf{r}$ is the radius of S, and where the terms in the expansion come from Eq. (A-36). These terms can be written.

$$d\mathbf{r}\cdot\nabla V = dx\,\frac{\partial V}{\partial x} + dy\,\frac{\partial V}{\partial y} + dz\,\frac{\partial V}{\partial z} \tag{A-51}$$

and $$\tfrac{1}{2}(d\mathbf{r}\cdot\nabla)^2 V = \tfrac{1}{2}(dx)^2\,\frac{\partial^2 V}{\partial x^2} + \cdots + dx\,dy\,\frac{\partial^2 V}{\partial x \partial y} \tag{A-52}$$

where the omitted terms in Eq. (A-52) are obvious. The average of Eq. (A-51) vanishes, since quantities like dx occur in pairs which cancel, and the average of $V(\mathbf{r})$ is identical to $V(\mathbf{r})$, since $\mathbf{r}$ is the coordinate of the center of the sphere. Finally,

$$< (d\mathbf{r})^2 > = 3 < (dx)^2 >$$

so that $$B(V) = \frac{1}{24\pi}\nabla^2 V \tag{A-53}$$

as we predicted.

It is useful to express the Laplacian in *spherical coordinates* r, θ, and φ, which are defined as follows:

$$\left.\begin{aligned} x &= r \sin\theta \cos\varphi \\ y &= r \sin\theta \sin\varphi \\ z &= r \cos\theta \end{aligned}\right\} \tag{A-54}$$

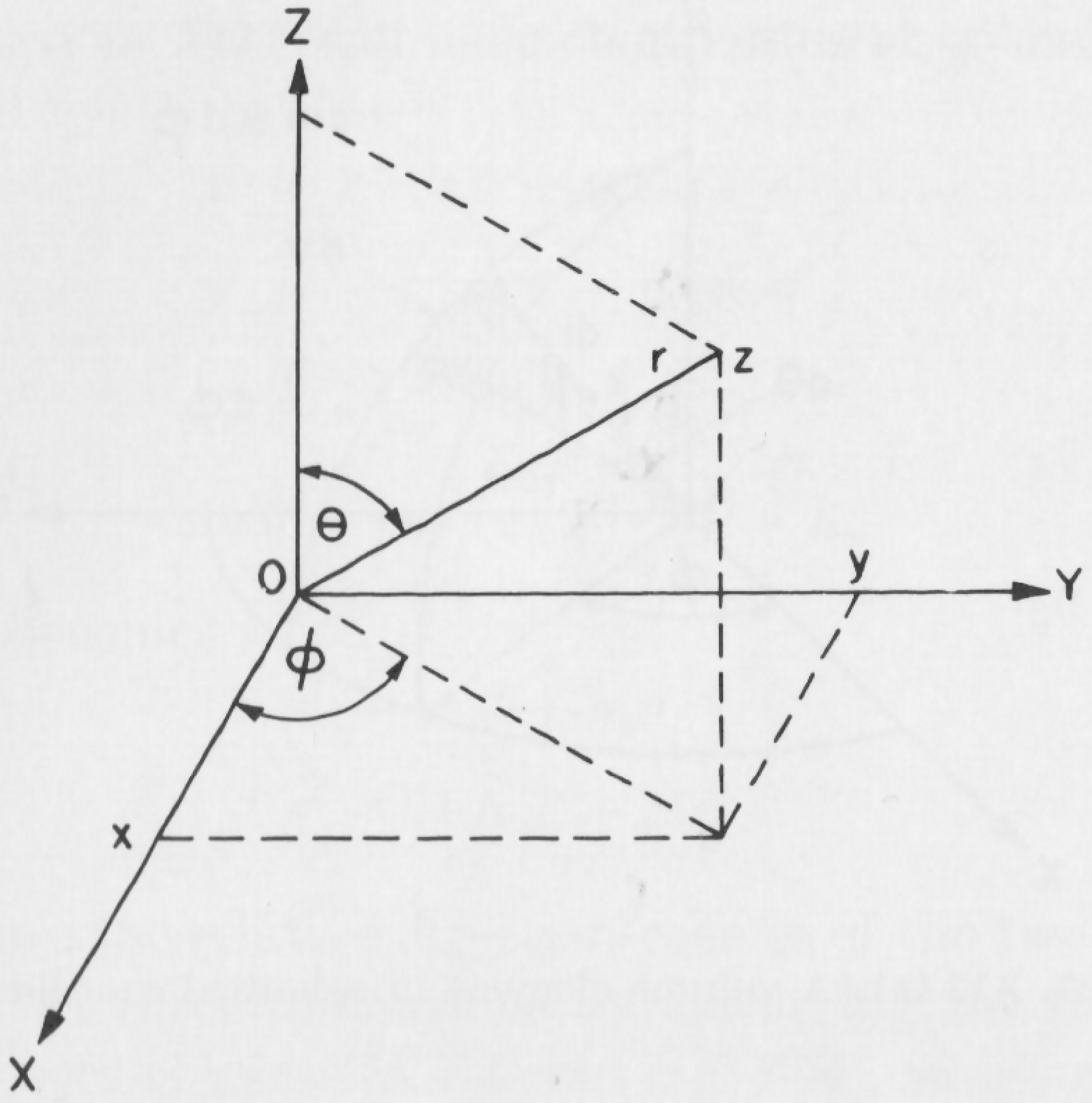


FIG. A13 (a). Spherical coordinates.

Figure A13a shows the basis of these definitions, which have the following inverses:

$$\left.\begin{aligned} r &= \sqrt{x^2 + y^2 + z^2} \\ \varphi &= \text{arc tan } y/x \\ \theta &= \text{arc cos } z/\sqrt{x^2 + y^2 + z^2} \end{aligned}\right\} \tag{A-55}$$

We note that when $\theta = 90$ deg, Eqs. (A-54) reduce to the well-known relations between rectangular and polar coordinates

$$x = r \cos\varphi$$

$$y = r \sin\varphi$$

By using relations of the form

$$\frac{\partial}{\partial x} = \frac{\partial r}{\partial x}\frac{\partial}{\partial r} + \frac{\partial \theta}{\partial x}\frac{\partial}{\partial \theta} + \frac{\partial \varphi}{\partial x}\frac{\partial}{\partial \varphi}$$

and $$\frac{\partial^2}{\partial x^2} = \frac{\partial^2 r}{\partial x^2}\frac{\partial^2}{\partial r^2} + \frac{\partial^2 \theta}{\partial x^2}\frac{\partial^2}{\partial \theta^2} + \frac{\partial^2 \varphi}{\partial x^2}\frac{\partial^2}{\partial \varphi^2}$$

$$+ 2\frac{\partial r}{\partial x}\frac{\partial \theta}{\partial x}\frac{\partial^2}{\partial r \partial \theta} + 2\frac{\partial \theta}{\partial x}\frac{\partial \varphi}{\partial x}\frac{\partial^2}{\partial \theta \partial \varphi} + 2\frac{\partial r}{\partial x}\frac{\partial \varphi}{\partial x}\frac{\partial^2}{\partial r \partial \varphi}$$

it can be shown (with a large amount of algebra) that the Laplacian becomes

$$\nabla^2 = \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial}{\partial r}\right) + \frac{1}{r^2 \sin\theta}\frac{\partial}{\partial \theta}\left(\sin\theta\frac{\partial}{\partial \theta}\right) + \frac{1}{r^2 \sin^2\theta}\frac{\partial}{\partial \varphi} \tag{A-56}$$

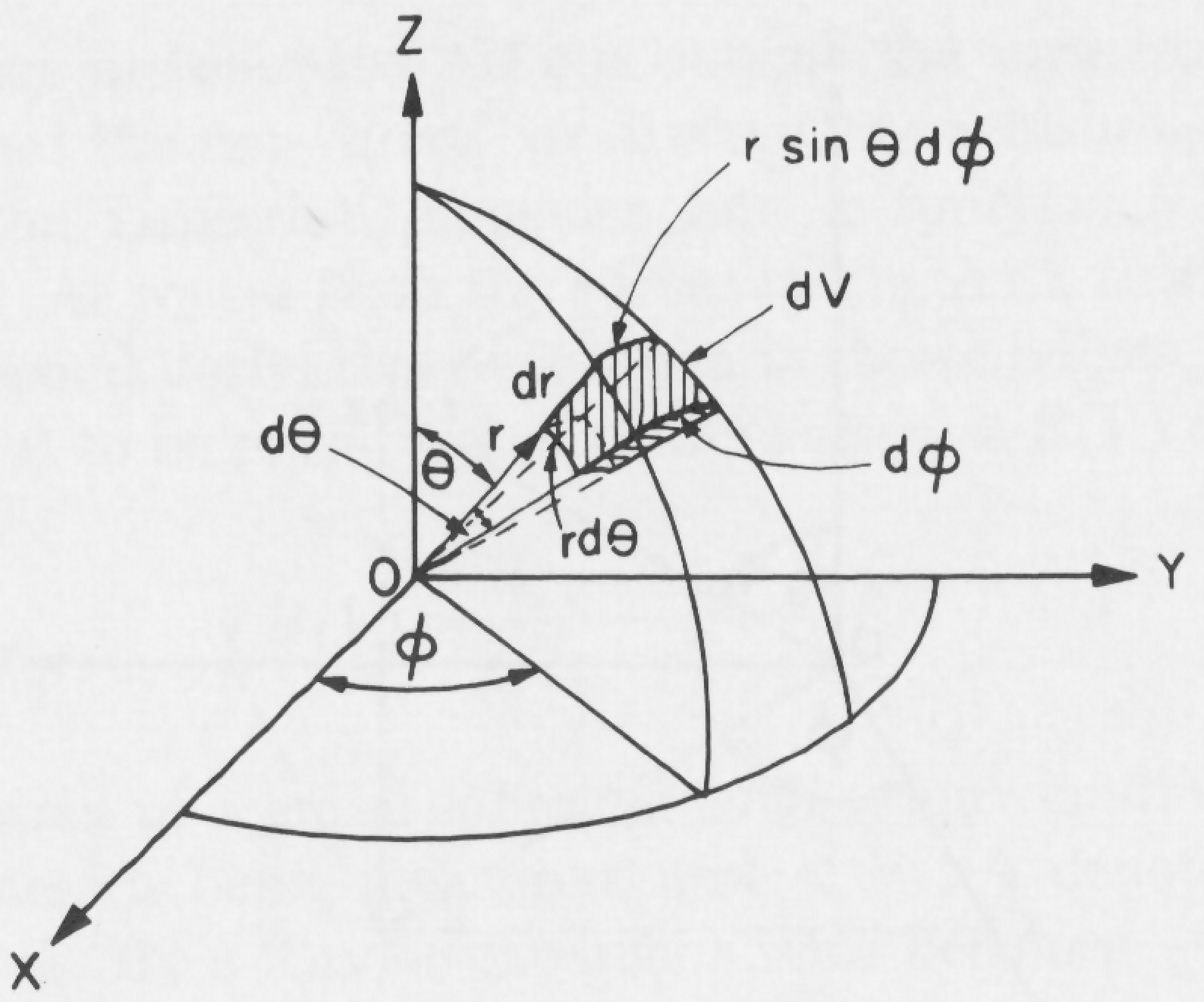


FIG. A13 (b). A volume element in spherical coordinates.

Figure A13b shows a *volume element* dV in spherical coordinates. The element is generated by the three orthogonal differentials dr, $d\theta$, and $d\varphi$, and is seen to be

$$\begin{aligned} dV &= dr(r\,d\theta)(r\sin\theta\,d\varphi) \\ &= r^2\sin\theta\,dr\,d\theta\,d\varphi \end{aligned}$$

A-4 Coordinate Transformations and Matrices

Consider a coordinate system XOY which is converted into a new system $X'OY'$ by rotation through an angle θ (Fig. A14). A point P which has coordinates x, y in XOY will have coordinates x', y' in $X'OY'$, the two sets being related by

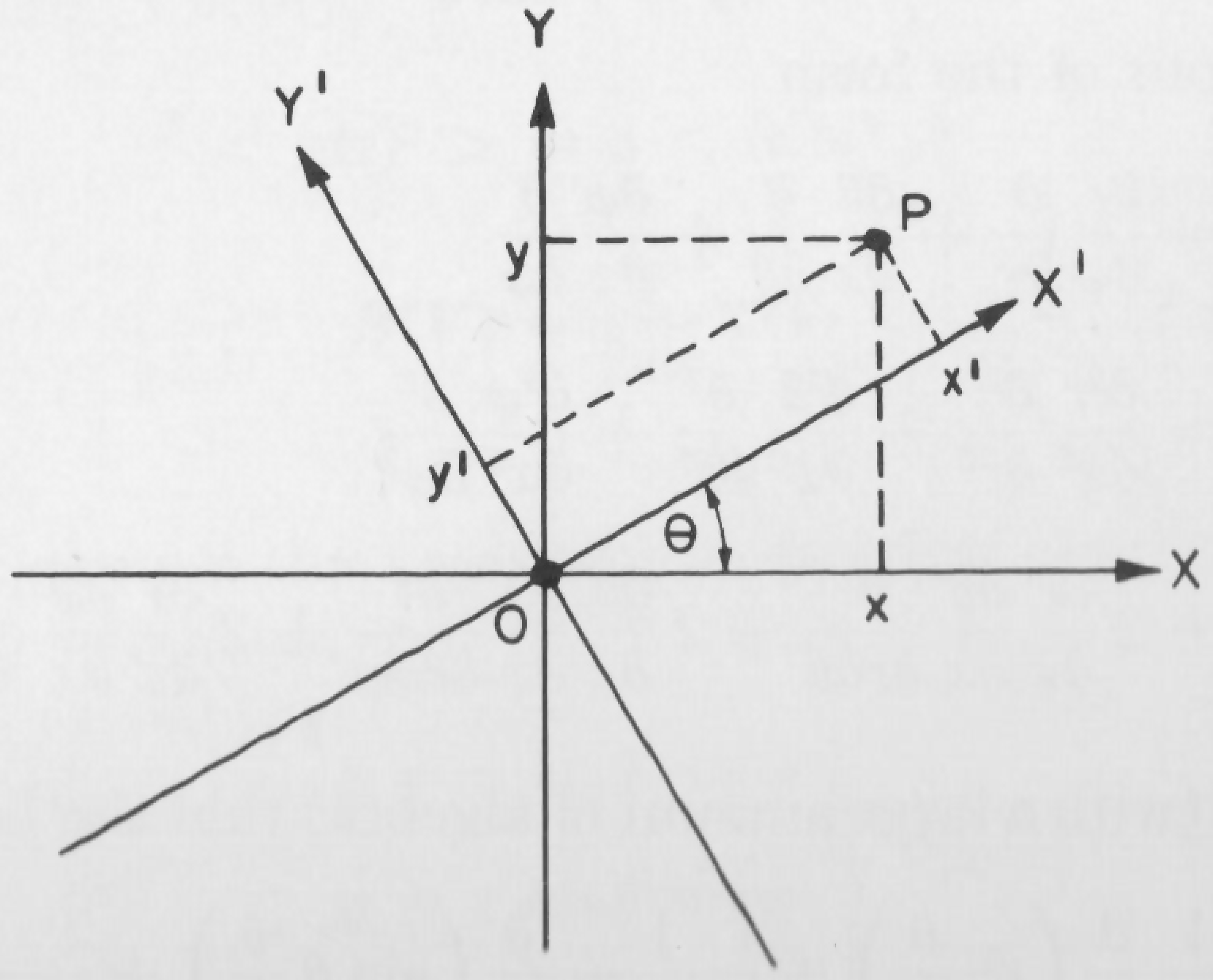


FIG. A14. Rotation of coordinates.

$$\left.\begin{aligned} x' &= x\cos\theta + y\sin\theta \\ y' &= -x\sin\theta + y\cos\theta \end{aligned}\right\} \tag{A-57}$$

If we let

$$\cos\theta = \alpha_x$$

$$\sin\theta = \cos(\pi/2 - \theta) = \alpha_y$$

then Eq. (A-57) becomes

$$\left.\begin{aligned} x' &= \alpha_x x + \alpha_y y \\ y' &= -\alpha_x x + \alpha_y y \end{aligned}\right\} \tag{A-58}$$

where α_x and α_y are the relative direction-cosines of the two systems. This can be extended to three dimensions by specifying the direction-cosines between the two sets of axes $OXYZ$ and $OX'Y'Z'$, which are

$$\begin{array}{c|ccc} & OX & OY & OZ \\ \hline OX' & \alpha_{11} & \alpha_{12} & \alpha_{13} \\ OY' & \alpha_{21} & \alpha_{22} & \alpha_{23} \\ OZ' & \alpha_{31} & \alpha_{32} & \alpha_{33} \end{array} \tag{A-59}$$

The equations corresponding to Eq. (A-58) are then

$$\left.\begin{aligned} x' &= \alpha_{11}x + \alpha_{12}y + \alpha_{13}z \\ y' &= \alpha_{21}x + \alpha_{22}y + \alpha_{23}z \\ z' &= \alpha_{31}x + \alpha_{32}y + \alpha_{33}z \end{aligned}\right\} \tag{A-60}$$

Consider now a second rotation specified by direction cosines $\alpha'_{i\,j}$ to produce a set of coordinates x'', y'', z''. Then

$$\left.\begin{aligned} x'' &= \alpha'_{1\,1}x' + \alpha'_{1\,2}y' + \alpha'_{1\,3}z' \\ &= \alpha'_{1\,1}(\alpha_{11}x + \alpha_{12}y + \alpha_{13}z) + \text{etc.} \\ &= (\alpha'_{1\,1}\alpha_{11} + \alpha'_{1\,2}\alpha_{21} + \alpha'_{1\,3}\alpha_{31})x + \text{etc.} \end{aligned}\right\} \tag{A-61}$$

with corresponding equations for y'' and z''. If we now let the set of coordinates x, y, z be converted directly to x'', y'', z'' by the group of direction-cosines $\alpha''_{i\,j}$, then Eq. (A-61) shows that

$$\begin{aligned} \alpha''_{i\,j} &= \alpha_{i1}\alpha'_{1\,j} + \alpha_{i2}\alpha'_{2\,j} + \alpha_{i3}\alpha'_{3\,j} \\ &= \sum_k \alpha_{ik}\alpha'_{k\,j} \end{aligned} \tag{A-62}$$

The arrangement [Eq. (A-59)] of the α_{ij} is called an *array,* and an array whose rule of multiplication corresponds to Eq. (A-62) is defined as a *matrix.* Let $\mathbf{A}$ denote the 3 × 3 matrix:

$$\mathbf{A} = \begin{bmatrix} \alpha_{11} & \alpha_{12} & \alpha_{13} \\ \alpha_{21} & \alpha_{22} & \omega_{23} \\ \alpha_{31} & \alpha_{32} & \alpha_{33} \end{bmatrix} \tag{A-63}$$

with similar definitions for $\mathbf{A}'$ and $\mathbf{A}''$. Also, define a *one-column matrix* as a 3 × 3 matrix with non-zero elements in only one column. Regard the set of coordinates x, y, z as a one-column matrix $\mathbf{r}$:

$$\mathbf{r} = \begin{bmatrix} x & 0 & 0 \\ y & 0 & 0 \\ z & 0 & 0 \end{bmatrix}$$

with a similar definition for $\mathbf{r}'$, corresponding to x', y', z'. Then the rule for matrix multiplication Eq. (A-62) shows that Eq. (A-60) can be written in matrix form as

$$\mathbf{r}' = \mathbf{A}\mathbf{r} \tag{A-64}$$

and Eq. (A-61) becomes

$$\mathbf{r}'' = \mathbf{A}'\mathbf{r}' = \mathbf{A}'\mathbf{A}\mathbf{r}$$
$$= \mathbf{A}''\mathbf{r}$$

or

$$\mathbf{A}'' = \mathbf{A}'\mathbf{A} \tag{A-65}$$

Hence, the rule for matrix multiplication provides for the matrix of two successive rotations to be the product of the individual rotation matrices.

The *inverse* $\mathbf{A}^{-1}$ of a matrix is defined by the relation

$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$$

where $\mathbf{I} = [\delta_{rs}]$ is called the *unit matrix.* In three dimensions, $\mathbf{I}$ has the form

$$\mathbf{I} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

It can be shown that

$$\mathbf{A}^{-1} = (A_{ji})/\det \mathbf{A}$$

where A_{ji} is the 2 × 2 determinant formed by striking out the row and column containing the element α_{ji} of $\mathbf{A}$ and multiplying by $(-1)^{i+j}$. A_{ji} is called the *co-factor* of α_{ij}, and det $\mathbf{A}$ denotes the determinant of $\mathbf{A}$. By direct expansion, we see that

$$\det \mathbf{A} = \alpha_{11}(\alpha_{22}\alpha_{33} - \alpha_{23}\alpha_{32}) + \text{(two similar terms)}$$
$$= \sum \alpha_{ik}A_{jk}$$

Then

$$\mathbf{A}\mathbf{A}^{-1} = \sum_k \alpha_{ik}\alpha^{-1}_{k\,j} = (\sum \alpha_{ik}A_{jk})/\det \mathbf{A} = \mathbf{I}$$

Note that $\mathbf{A}^{-1}$ exists only if det $\mathbf{A} \neq 0$; such a matrix is said to be *non-singular.*

A-5 Introduction to Tensors

Figure A15a shows two identical springs lying along the coordinate axes. They are joined at their centers and fastened down at the ends. If we tie a string to this junction and pull slightly, the displacement $\mathbf{r}$ of the common center lies along the direction of the restoring force $\mathbf{F}$. We can see this by realizing that, for a small displacement, a force along the x-axis produces a displacement given by

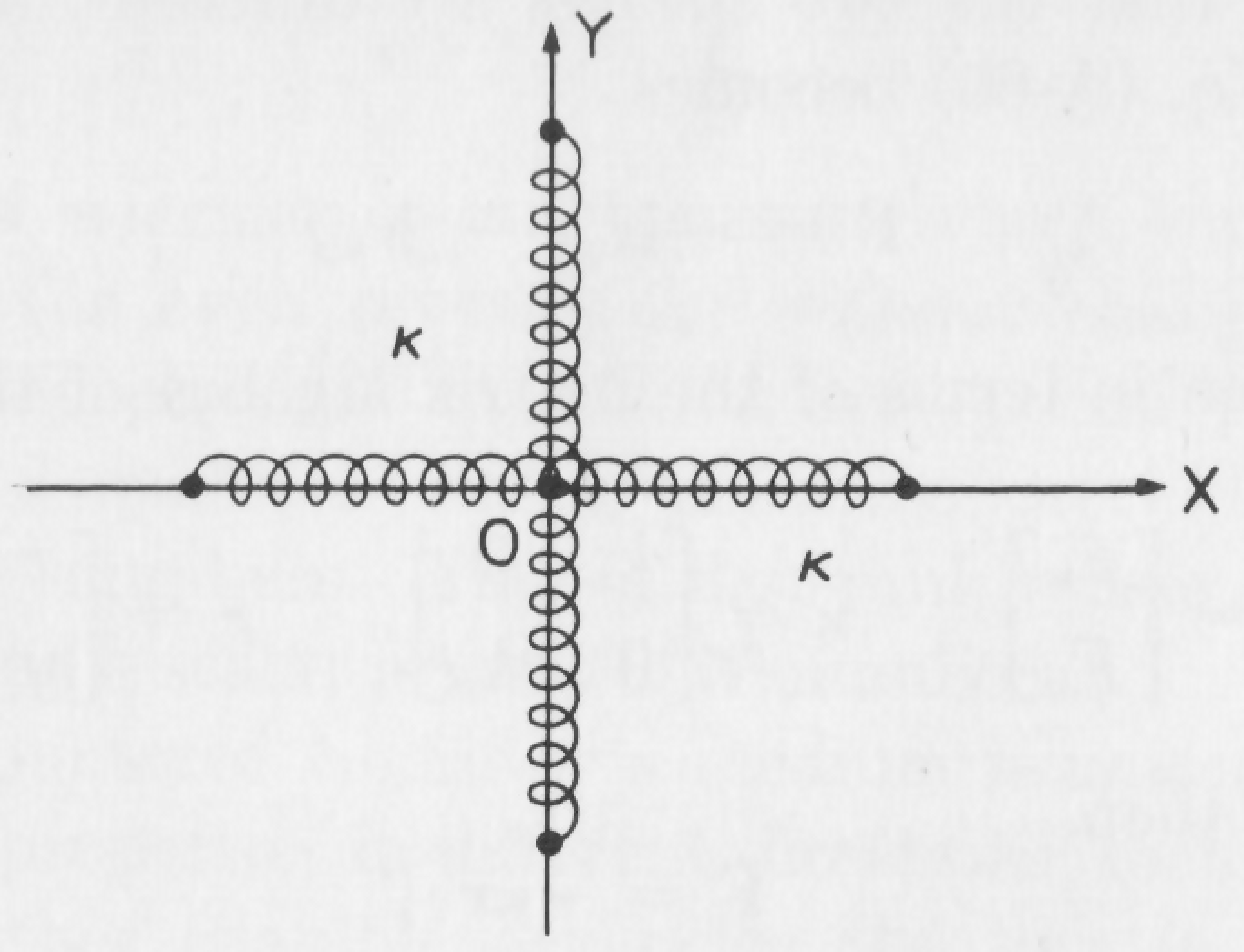


FIG. A15(a). An interpretation of a tensor. (See text.)

$$F_x = -\kappa x$$

and, similarly,

$$F_y = -\kappa y$$

where κ is the Hooke's law constant of both springs. For an arbitrary force, then,

$$\mathbf{F} = \mathbf{i}F_x + \mathbf{j}F_y = -\kappa(\mathbf{i}x + \mathbf{j}y) = -\kappa\mathbf{r} \tag{A-66}$$

and **F** and **r** are colinear (Fig. A15b). The approximation of small displacements is necessary, so that the y-spring does not affect the x-motion, and vice versa.

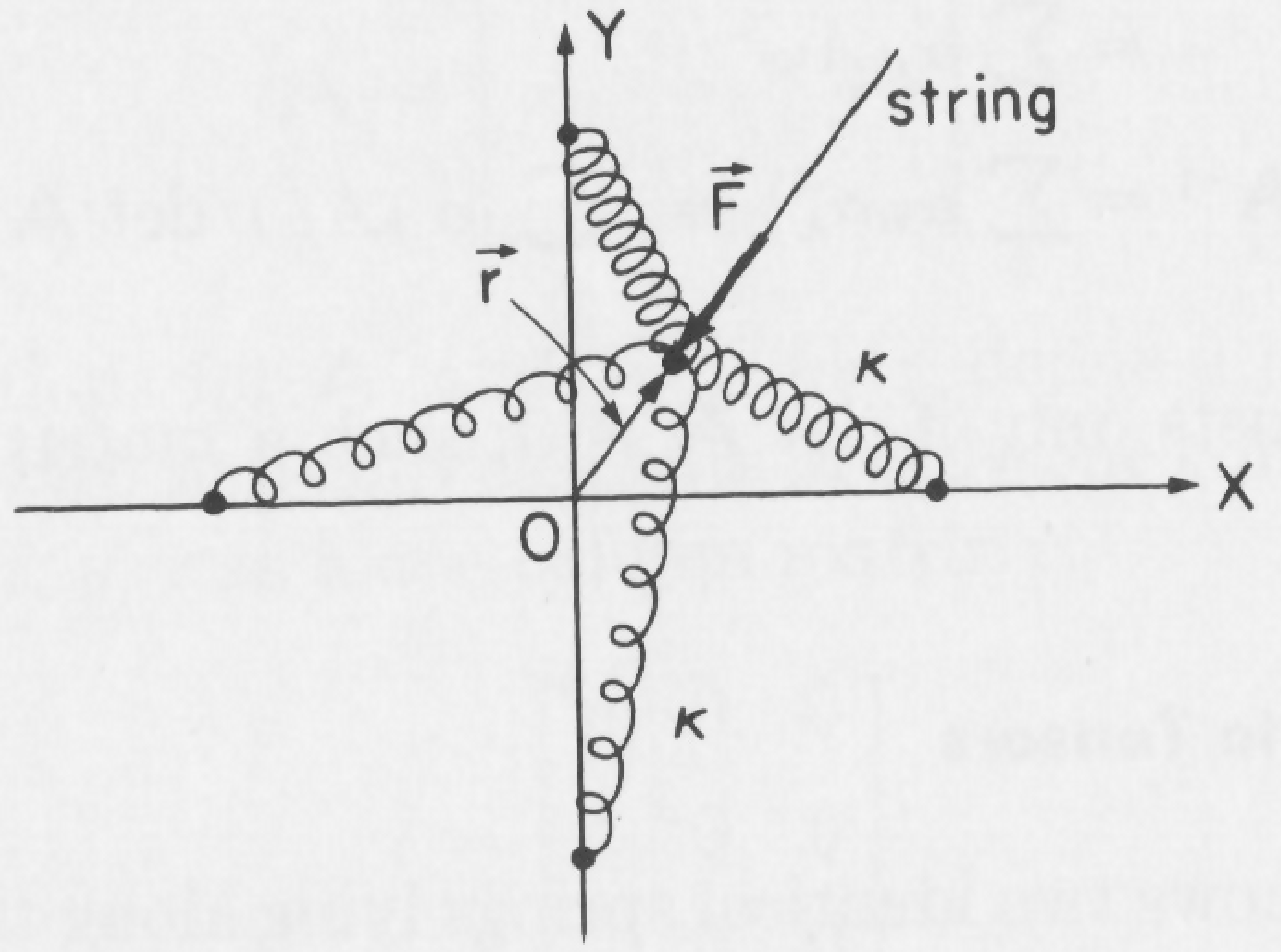


FIG. A-15(b).

Now, suppose that the two springs are different, and have constants κ_x and κ_y. Then Eq. (A-66) becomes

$$\mathbf{F} = -\mathbf{i}\kappa_x x - \mathbf{j}\kappa_y y \tag{A-67}$$

This can be written in terms of the matrix algebra of the previous section if we define

$$\mathbf{F} = \begin{bmatrix} F_x \\ F_y \end{bmatrix} \qquad \boldsymbol{\kappa} = \begin{bmatrix} k_x & 0 \\ 0 & k_y \end{bmatrix} \qquad \mathbf{r} = \begin{bmatrix} x \\ y \end{bmatrix}$$

and Eq. (A-67) is then

$$\mathbf{F} = -\boldsymbol{\kappa}\mathbf{r} \tag{A-68}$$

where κ is called a *tensor*. In three dimensions, Eq. (A-68) has the same form, provided that we define

$$\mathbf{F} = \begin{bmatrix} F_x \\ F_y \\ F_z \end{bmatrix} \qquad \boldsymbol{\kappa} = \begin{bmatrix} k_x & 0 & 0 \\ 0 & k_y & 0 \\ 0 & 0 & k_z \end{bmatrix} \qquad \mathbf{r} = \begin{bmatrix} x \\ y \\ z \end{bmatrix} \tag{A-69}$$

In this book, we shall designate tensors by bold-face Greek symbols; in written notes, a convenient symbol is the double arrow: $\overset{\rightrightarrows}{\kappa}$.

To examine the physical meaning of tensors, consider Fig. A15c, which shows the situation when the two springs are different. The applied force **F** has resulted in a displacement **r**, which is, in general, in a different direction from **F**, as required by Eq. (A-67). Hence, we can think of a tensor as a "machine" into which we insert a vector **F** and out of which comes another vector **r** which is different in all respects: magnitude, direction, and units. Under special circumstances, the tensor $\boldsymbol{\kappa}$ will become a scalar; that is, when $\kappa_x = \kappa_y = \kappa$. For then

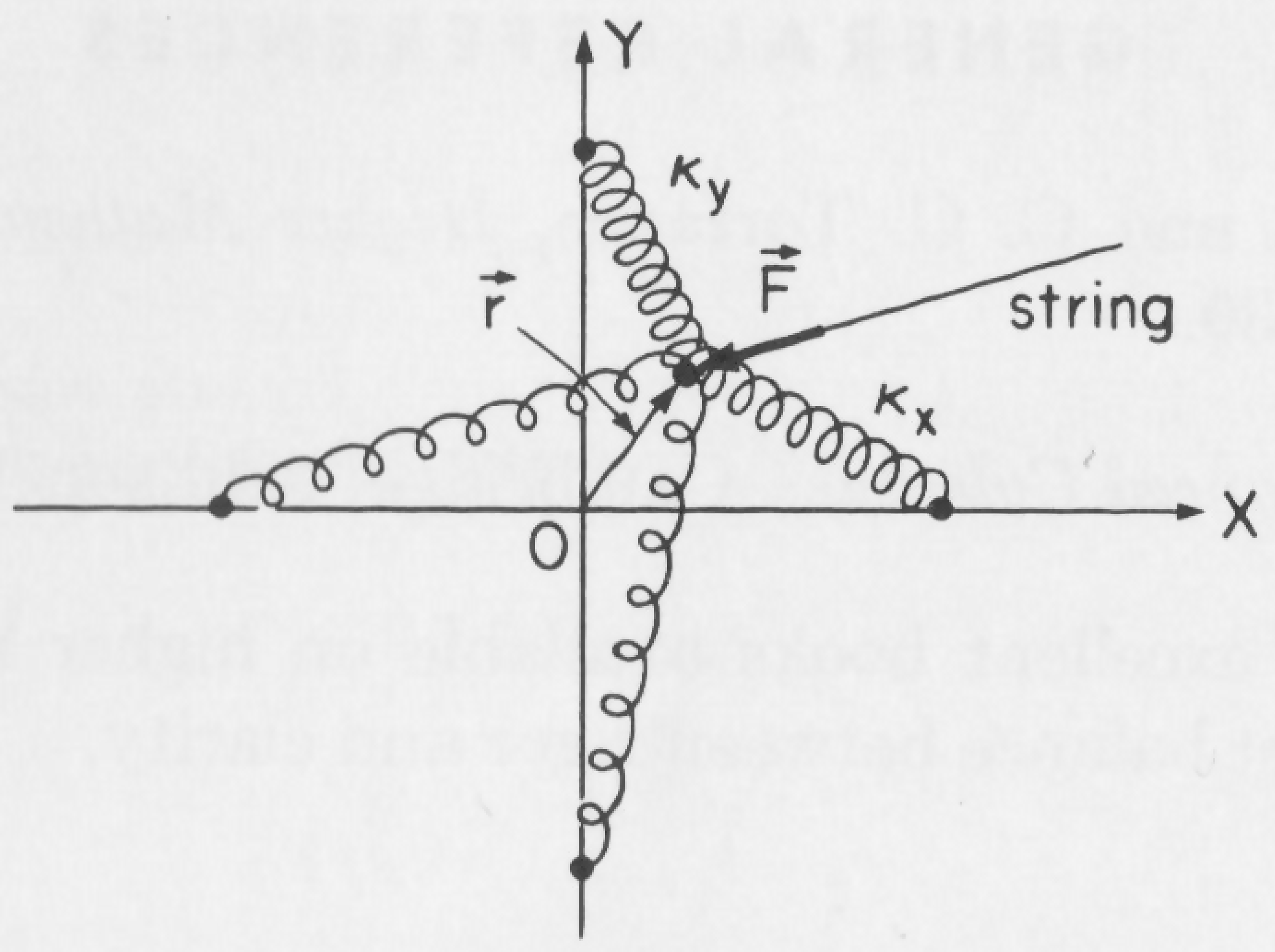


FIG. A-15(c).

$$\mathbf{F} = -\begin{pmatrix} \kappa & 0 \\ 0 & \kappa \end{pmatrix}\mathbf{r} = -\kappa\mathbf{r}$$

Hence, we think of a scalar as an operator which simply stretches or contracts a vector (or even reverses it) without changing its direction angles. Actually, all physical quantities are special forms of tensors. In three-dimensional space, a tensor (or more precisely, a *second-rank* tensor) is a set of 3^2 numbers. Then a first-rank tensor, or a vector, is a set of 3^1 numbers, and a scalar is a set of 3^0 numbers.

Tensors are encountered wherever a medium is *anisotropic;* i.e., whenever it has different properties in different directions, such as for the springs of Fig. A15c. Another example occurs in the text in connection with Ohm's law, which is written

$$\mathbf{J} = \sigma\boldsymbol{\mathcal{E}}$$

in Chapter 1, since **J** and $\boldsymbol{\mathcal{E}}$ are colinear. However, in Chapter 7 we show that the application of a magnetic field **B** converts this into

$$\mathbf{J} = \boldsymbol{\sigma}\boldsymbol{\mathcal{E}}$$

where the conductivity is now a tensor:

$$\boldsymbol{\sigma} = \begin{bmatrix} \sigma_{xx} & \sigma_{xy} & \sigma_{xz} \\ \sigma_{yx} & \sigma_{yy} & \sigma_{yz} \\ \sigma_{zx} & \sigma_{zy} & \sigma_{zz} \end{bmatrix}$$

and **J** and $\boldsymbol{\mathcal{E}}$ are no longer colinear.

GENERAL REFERENCES

Burington, R. S. and C. C. Torrance, *Higher Mathematics.* New York: McGraw-Hill, 1939.

Kaplan, W., *Advanced Calculus.* Cambridge: Addison-Wesley, 1952.

Of the many excellent books available on higher mathematics, these two strike the best balance between rigor and clarity.

INDEX

Absolute thermopower (*See* Thermopower)
Acceleration, 2
Acceptor atom, 80
Accumulation, 127
Ampere, 6
Amplification (*See* Transistor)
Angstrom unit, 2
Atmosphere (as a pressure unit), 3
Atom, hydrogen, 29 (*See also* Schroedinger equation)
Avogadro's number, 60
Axis, alternating, 175
Axis, screw, 177

Bands, energy (*See* Energy bands)
Barrier, space-charge, 92–93, 131
  generation-recombination in, 137–148
Beats, 15
Bias voltage, 104
Bloch waves, 39
Boltzmann's constant, 60, 63–64
Boltzmann transport equation, 230–232
  for ellipsoidal surfaces, 233–240
  iterative solution, 253
  in a magnetic field, 236–240
  for warped surfaces, 253–262
Box, potential, 26–28
Bragg–Laue diffraction, 184–189
Bravais lattices, 163–170
Brillouin zones, 183–190
  first, of reciprocal lattice, 189
  reduced zone of, 190
  for silicon and germanium, 204–206
  for tellurium, 264
Built-in potential, 93
Bulginess, 326
Bulk modulus, 305

Capacitance, 5
  of junction, 103
Capacitor, 5
Capture rate, 141
Carrier density, 81
  calculation of, 81–84
  intrinsic, 84
  relation to chemical potential, 85
Carrier exclusion, 127
Centripetal force, 207
Charge, 1
  of electron, 4
  quantity of, 4
Chemical equilibrium, 48
Chemical potential, 46
  for electrons and holes separately, 105
  interpretation, 87, 93
  relation to carrier densities, 85
Coefficients, reflection and transmission, 304
Collision time, 225
Collisions, electron-phonon, 303
Compensation, 281–283
Concentration of impurities:
  determination from Hall coefficient, 283–287
  determination from resistivity, 277–282
Conductivity, 6
  in terms of average collision time, 225–226
Conductivity tensor, 239
  for N-type germanium, 244–245
Conjugate forces and fluxes, 215–218
Contacts, infinite recombination rate, 111
Continuity equation, 109, 124–125
Coordinates, spherical, 326–327
Coordinate transformations, 328–331
Coulomb's law, 4

Crystal:
  covalent, 76
  definition, 38
  symmetry properties, 151–183
  (*See also* Lattices)
Curl, 321
Current, electric, 6
Current density, 6, 212–213
  from Boltzmann's equation, 233–235
Cyclic boundary conditions, 71
Cyclotron resonance, 206–210

de Broglie's hypothesis, 17
Debye length, 97
Degeneracy, removal of, in energy bands, 202, 204
Degeneracy, spin, 283–284, 286–287
Degenerate semiconductor, 83
Del, 320
Density, electric current (*See* Current density)
Depletion layer (*See* Barrier, space-charge)
Derived units, 2
Detailed balance, 140–141
Dielectric constant (*See* Permittivity)
Diffusion, 86
Diffusion constants, 86
Diffusion length, 111
  downstream, 127
  high level, 131
Diode, PNN^-, 122–123
Diode, *PN* (*See PN*-junction)
Disorder, 51, 58, 61
Distributions, statistical:
  approximation of Fermi-Dirac by Maxwell-Boltzmann, 83, 87–88
  Bose-Einstein, 56–60
  Fermi-Dirac, 56–60
  Maxwell-Boltzmann, 56–60
Divergence, 320
Divergence theorem, 325
Donor atom, 77
Dopes, 80
Drift length, 144

Effective mass (*See* Mass, effective)
Einstein equation:
  for mass-energy equivalence, 17
  for mobility-diffusion coefficient relation, 86
  for photoelectric effect, 15
Electrochemical potential:
  definition, 67

Electrochemical potential: (*Cont.*)
  and Fermi level, 68
  and generalized Ohm's law, 85–86
  and high-level injection, 131
  identification with applied voltage, 107
  interpretation, 68–69
  separate definitions for electrons and holes, 105
Electrons:
  bound, 81
  excess, 77
  free, 41, 81
Electron shell, 35–36
Electron spin, 35
Electron-volt, 5
Energy, 2
  free, 53
  internal, 45
  kinetic, 2
  potential, 2
Energy bands:
  conduction, 74
  dependence on lattice constant, 43
  forbidden, 74
  as a function of distance, 77
  occupancy in insulators and conductors, 74
  origin of allowed and forbidden, 40
  for *PN*-junction in equilibrium, 92–93, 106–109
  for *PN*-junction under low-level injection, 129
  for silicon and germanium, 203–210
  for tellurium, 263, 271–273
  valence, 74
Energy current density, 216, 218–219
Energy surfaces:
  ellipsoidal, 206
    in many-valley semiconductors, 242–243
  warped, 209
    in P-type silicon and germanium, 255–259
Enthalpy: 52
Entropy:
  and available work, 50
  calculation of, 49–50
  definition, 46
  and disorder, 51
  extremum condition for, 60
  volume rate of generation, 216–217
Entropy current density, 216, 219
Equations, homogeneous, 39

Equilibrium:
  chemical, 48
  mechanical, 47
  thermal, 47
Ettingshausen coefficient, 274
Excitation:
  optical, 75
  thermal, 75
Exclusion (*See* Carrier exclusion)
Exclusion principle, 35
Exhaustion of impurities, 285
Extensive variables, 45
Extraction, 127

Farad, 5
Fermi function:
  definition, 64
  effect on occupancy of levels, 88–89
  interpretation, 65
  relation to band structure, 86–90
  temperature dependence, 89–90
Fermi level:
  definition, 64
  effect on occupancy of levels, 88–89
  interpretation, 65–67
  quasi-Fermi level, 106
  relation to electrochemical potential, 68
  relation of position to impurity concentration, 90
  temperature dependence, 89–90
  for traps, 143, 146
  (*See also* Thermopower)
Fick's law, 86
Field:
  electric, 5
  magnetic, 7
Figure of merit, thermoelectric, 294–295
First law of thermodynamics, 45
Flux-density, 7
Force, 2
Force, harmonic, 8
Fourier's law, 212
Free energy, 53
Frequency, 8–9
Fundamental quantities, 1

Gain (*See* Transistor)
Galvanomagnetic coefficients:
  definitions and interpretation, 223–229
  for *N*-type silicon and germanium, 242–252
  for *P*-type germanium, 259–260

Galvanomagnetic coefficients: (*Cont.*)
  for *P*-type silicon, 260
  for tellurium, 263–273
Generation-recombination processes, 137–148
  effect on *PN*-junctions, 143–148
Generation-recombination current, 145–148
Gibbs function, 52–53
Glide, diagonal, 180
Glide line, 159, 162
Glide, plane, 177
Gradient, 320
  interpretation of, 322–323
Ground state, 35
Groups, space, 177–183
Groups, symmetry, 156–183
  dihedral, 173
  holosymmetric, 181
  plane point, 155–157, 158–163
  three-dimensional point, 170–177

Hall angle, 228
Hall coefficient and effect:
  in anisotropic semiconductors, 240
  description, 225–226
  high-field, 248
  low-field, 247
  in *N*-type silicon and germanium, 247–251
  in *P*-type silicon and germanium, 259–260
  in tellurium, 263–273
Hall-Shockley-Read recombination, 137–142
Heat, 45
Heat content, 45, 54
Heat current density, 235
Helmholtz function, 52–53
Hermann-Mauguin crystal classification, 170–173
Hermite equation, 21
Hole conduction, 78–81
Hooke's law, 8
Hydrogen atom, 29–34
Hyperbolic functions, 40
Hyperspheres, 61–63

Injection, 104–105
  high-level, 129–137
Injection efficiency, 112–113
Integrals, line, surface, and volume, 323–325

Intensive variables, 45
Internal energy, 45
International crystal classification (*See* Hermann-Mauguin)
Interstitial atoms, 137
Intrinsic carrier density, 84
Ionized impurities, 88
Irreversible process, 48
Irreversible thermodynamics, 212–222

Joule, 2
Junction (*See PN*-junction)

Kelvin relations, 220–221
Kilogram, 2
Kronecker delta, 319

Lagrange's multipliers, 59
Laguerre equation, associated, 31
Laplacian, 320
  interpretation, 326
  in spherical coordinates, 326–327
Lattices, 153–183
  Bragg-Laue reflection in reciprocal, 188
  Bravais, 163–170
  cubic, 166
  diamond, 179–180
  direct, 187
  hexagonal, 155–156
  oblique, 153–154
  orthorhombic, 165
  polonium, 182
  plane, 158–163
  reciprocal, 187–188
  rectangular, 155–156
  square, 155, 162
  table of, 167
  tellurium, 181–183
  tetragonal, 166
  trigonal, 166
Lattice constant, 38
LCAO method (*See* Tight-binding method)
Leakage current (*See* Reverse saturation current)
Legendre equation, associated, 30
Legendre transform, 51
Length, 1
Levels, energy:
  free electron, 41
  harmonic oscillator, 22
  hydrogen atom, 31
  impurity, 77

Levels, energy: (*Cont.*)
  intrinsic, 84
  potential box, 28
  potential well, 26
  relation to energy states, 73
Lifetime (*See* Minority carrier lifetime)

Magnetoresistance, 228–229
  in silicon and germanium, 252, 262
  in tellurium, 269–273
Mass, 1
Mass, effective, 42
  anisotropy, of, 208
  anisotropy ratio, 239
    for N-type germanium, 250
    for N-type silicon, 250
  density-of-states, 286
  effect on impurity levels, 78
  heavy and light, 210
  longitudinal, 206
  tensor, 237
  transverse, 206
Mass, relativistic change in, 17
Matrix algebra, 328–331
Maxwell's relations, 53, 55
Metals, 75
  thermopower of, 294
Meter, 2
Mean free path, 246
Mean free time between collisions, 299
Micron, 2
Microsecond, 2
Miller indices, 185–186
Minority carrier lifetime, 104, 108, 125
  dependence on concentration, 138–142
  limiting values, 138
  measurement of, 308–310
MKS units, 2, 4
Mobility, 7
  drift, 225
  Hall, 228, 248
  lattice, 305–306
  magnetoresistance, 229
  and scattering, 296–308
  temperature variation, 249
  (*See also* Scattering)
Momentum, 2
  angular, 3

Nernst coefficient, 274
Newton (as force unit), 2
Normalization constant, 32, 34
N-type semiconductor, 80

Ohm, 6
Ohm's law, 6, 212
  and electrochemical potential, 68–69, 85
  generalized, 219
Onsager principle, 212–215
  applications, 217–221, 268, 274
OPW method, 203
Orbitals, atomic (*See* Tight-binding method)
Orthogonal functions, 200
Oscillator, quantum, 21–23

Particle current density, 216
Pauli exclusion principle, 35, 65, 74
Peltier coefficient and effect, 220–221
Perfect gas, 60–63
Periodic table, 37
Permittivity, 4–5
Permutation symbols, 320
Phase, 9
Phase space, 61
Photoconductive decay, 308–309
Photoelectric effect, 11–12, 15
Photon, 13
  group and phase velocity of, 16
Planar Hall coefficient:
  applications, 241–242
  definition, 240
  in germanium, 252
Planck's constant, 14
PN-junction:
  abrupt, 100
  bias polarities, 104
  capacity, 103
  currents in, 112
  definition, 91
  in equilibrium, 94–102
  fabrication of, 91–92
  formation of barrier layer, 92
  high-level, 129–137
  injection, 104–105
  low-level, 103–113
  rectifying action, 103–104
  symmetric, 100
Poisson's equation, 94–97, 109
Potential, electrostatic, 5, 67
Potential, periodic, and energy bands, 36–43, 71–74
Power, 2, 7
Pressure, 3
  thermostatic definition, 46
Primitive cell, 154
Primitive vectors, 154–155

Probability and wave functions, 23
Probability-density, 23, 30, 31
Propagation constant, 9
P-type semiconductor, 80

Quantum, 13
Quantum numbers:
  harmonic oscillator, 22
  hydrogen atom, 30–31
  potential box, 27
  potential well, 26
  principal, 31
Quantum theory, 13
Quasi-Fermi level, 106

Radian, 3–4
  spherical, 95
Recombination centers, 137
Recombination current, 125, 135
Rectifiers, 104
Reduced zone (*See* Brillouin zones)
Relative thermopower (*See* Thermopower)
Relaxation time:
  in Boltzmann equation, 231
  effect on thermopower, 293–294
  as a tensor, 236
Reservoir, heat, 49
Resistance, 6
Resistivity, 6
  relation to concentration, 277–282
Resistivity tensor, 241
Rest-energy, 18
Rest-mass, 17
Reverse saturation current, 104, 112
Reversible process, 49
Righi-Leduc coefficient, 274

Scalar product, 317
Scalar quantity, 315
Scattering:
  accoustical lattice, 306–307
  Brooks-Herring, 301
  Conwell-Weisskopf, 000
  ionized impurity, 296–302
  lattice, 302–308
  and mobility, 296–308
  Rutherford, 296–298
  of X-rays, 184
Schoenflies crystal classification, 170–173
Schroedinger equation, 19–20
  harmonic oscillator, 21
  hydrogen atom, 29
  periodic potential, 38

- Schroedinger equation: (*Cont.*)
  - potential box, 26
  - rectangular well, 24
- Second, 2
- Secular equation, 199–201
- Seebeck effect (*See* Thermopower)
- Semiconductor:
  - degenerate, 83
  - impurity, 75
  - intrinsic, 75
- Shell, electron, 35
- Solid angle, 94
- Soret coefficient, 288
- Space-charge layer, 93
- Space-charge neutrality, 110, 130
- Specific heat capacity, 99
- Speed, 2
- Spin, 35
- Spin-orbit coupling, 204
- State:
  - degenerate, 73
  - electron, 28
- States, density of, 28, 60–61
  - effective, 84
- Statistics of semiconductors, 80–84
- Surface integral, 324
- Surface recombination velocity, 125–126
- Symmetry:
  - bilateral, 151
  - crystal, 151–183 (*See also* Lattices)
  - rotational, 151–152
  - translational, 151–152
- Systems, crystal, 166–170
  - minimum symmetry properties, 176

- Taylor series, 313–315
- Tellurium, 262–273
- Temperature, 1, 46
- Tensors, 331–333
- Thermal conductivity, 212–213
- Thermal current density, 212–213
- Thermoelectric power (*See* Thermopower)
- Thermomagnetic effects, 273–275
- Thermopower:
  - comparison of metals and semiconductors, 290–295
  - definition, 219
  - determination of majority carrier type, 287–288
  - graphical interpretation, 292–295
  - relation to Fermi level, 291–294
- Thermostatics, 44–55
- Thompson effect, 220–221
- Tight-binding method, 190–198
  - for tellurium, 199–203
- Time, 1
- Transistor:
  - amplification, low-level (alpha), 116
    - variation with emitter current, 136 144-146
    - variation with generation-recombination current, 148
  - base transport factor (beta), 116
  - calculation of electron and hole currents, 115–117
  - characteristic curves, 119–120
  - common emitter and common base configurations, 118
  - distribution of currents, 113–114
  - enhancement of emitter efficiency, 119
  - high-level, 129–137
  - measurement of basewidth, 120–121
  - *PNP*, 113–122
  - PNN^-, 122–128
  - structure, 113–114
- Transport integrals, 235, 240, 247, 265, 291
- Trapping, 137
  - effect on lifetime, 309–310

- Undetermined multipliers, 58–59
- Units, MKS, 2

- Valleys, energy-band, 206
- VATVUS diagram, 54
- Vectors, 315–328
  - curl, 321
  - definition, 315
  - divergence, 320, 324–325
  - gradient, 320–322
  - Laplacian, 320, 326–327
  - scalar product, 317
  - sum, 316
  - unit, 316
  - vector product, 317–319
- Velocity, 2
  - angular, 3
  - group, 16
  - phase, 16
- Voltage, 5

- Watt, 2
- Wave equation, classical, 11
- Wavelength, 9
- Wave-particle duality, 17
- Weber (magnetic unit), 7
- Work, 2